PROPERTIES OF PREDICTORS IN OVERDIFFERENCED NEARLY NONSTATIONARY AUTOREGRESSION

Ismael Sánchez and
Daniel Peña

95-58

WORKING PAPERS

Universidad Carlos III de Madrid

Working Paper 95-58
Statistics and Econometrics Series 24
December 1995

Departamento de Estadística y Econometría
Universidad Carlos III de Madrid
Calle Madrid, 126
28903 Getafe (Spain)
Fax (341) 624-9849

PROPERTIES OF PREDICTORS IN OVERDIFFERENCED NEARLY NONSTATIONARY AUTOREGRESSION

Ismael Sánchez and Daniel Peña*

Abstract

This paper analyzes the effect of overdifferencing a stationary AR($p+1$) process whose largest root is near unity. It is found that if the largest root is $\rho = \exp(-c/T^{\beta})$, $\beta > 1$, with T being the sample size and c a fixed constant, the estimators of the overdifferenced model ARIMA($p,1,0$) are root-T consistent. It is also found that this misspecified ARIMA($p,1,0$) has lower predictive mean square error than the properly specified AR($p+1$) model due to its parsimony. The consequences of this result are: (i) for forecasting purposes it is better to overdifferentiate than to underdifferentiate, (ii) the superiority of the overdifferenced predictor is small in the short term forecast but increases with the horizon, (iii) model selection based on predictive performance can lead to the wrong model in nearly nonstationary autoregression.

Key Words:
Autoregressive processes, near nonstationarity, overdifferencing, parsimony, predictive mean square error, unit roots.

*Departamento de Estadística y Econometría, Universidad Carlos III de Madrid. Work supported by the Spanish DGICYT grant PB93-0232.

1 Introduction

In this paper we investigate the consequences in estimation and prediction of overdifferencing an AR($p+1$) process whose largest root is inside the unit circle but close to one. Differencing is normally used to transform a homogeneous linear nonstationary time series into an stationary process, that is often modelled as an ARMA(p,q) process. Then it is said that the original series follows an ARIMA(p,d,q) process, where d is the number of differences required in order to obtain stationarity. We assume that d is an integer equal to the number of unit roots in the characteristic equation. When an autorregresive time series has its largest characteristic root close to the unit circle is said to be nearly nonstationary or near integrated. Given a small or moderate sample of this process, with largest root less than unity, it is very likely to conclude, due to the low power of unit roots tests in this case, that a difference should be applied. The differenced series will be noninvertible and is said to be overdifferenced.

Since the work of Fuller (1976) and Dickey & Fuller (1979) there has been a vast literature concerning the detection of unit roots in autoregressive polynomials. Also much attention has recently been paid to moving average (MA) unit roots testing (Tanaka 1990, Saikkonen & Luukkonen 1993, Tsay 1993) . However relatively little has been written on the consequences of a wrong detection. Previous work on the effect of overdifferencing can be found in Plosser & Schwert (1977, 1978) and Harvey (1981). Plosser & Schwert (1977) examine, using Monte Carlo techniques, the effect of overdifferencing in two examples: processes with a deterministic linear trend and stochastic regression models. They conclude that, in these situations, the loss in efficiency of both parameter estimators and prediction is not substantial, provided an MA parameter is included. Harvey (1981) proposes a finite sample predictor, based on the Kalman filter, for computing optimal predictions overcoming the problem of dealing with a noninvertible process. He also concludes that overdifferencing does not need to have serious implications for prediction provided a finite sample prediction procedure is used and an MA parameter is included. In this paper, we assume that the largest root of the AR polynomial is close to unity and, therefore, we will adopt as overdifferenced predictor the ARIMA($p,1,0$) model, where no MA component is involved. We will analyze the properties of the estimators of this ARIMA($p,1,0$) model and compare its predictive mean square error (PMSE) with the estimators of the properly specified AR($p+1$) model.

The effect of misspecification on the analitycal expression of PMSE has received much interest (Berk 1974, Bhansali 1978,1981, Davies & Newbold 1980, Tanaka & Maekawa 1984, Kunitomo & Yamamoto 1985 among others). Kunitomo & Yamamoto (1985) find a general expression for the PMSE of autoregressive processes of order m (m can be infinite) when a finite autoregression of order p is fitted (p can be larger, equal or smaller than m). In contrast with the approach developed here all of these authors assume that both the misspecified and the properly specified model are of the same order of differencing.

Misspecification in statistical model building is specially important when the correct model and the misspecified one are conceptually very different, as in the unit roots problem. Nevertheless in this article we prove that the PMSE of the overdifferenced model ARIMA$(p,1,0)$ is lower than the PMSE of the correct model AR$(p+1)$ if $\rho = \exp(-c/T^{\beta})$; $\beta > 1$, due to its parsimony. Some consequences of this result are:

1. For forecasting purposes it is better to overdifferentiate than to underdifferentiate. Therefore the low power of stationarity tests in autoregression is not as important in forecasting as in model identification.

2. The superiority of the overdifferenced predictor is small in the short term forecast but increases with the horizon.

3. Model selection based on predictive performance can lead to the wrong model in nearly nonstationary autoregression.

This paper is organized as follows. Section 2 introduces the model and notation. The consequences of overdifferencing in estimation are analyzed in section 3 and the effect on the PMSE for each predictor in section 4. Section 5 compares the PMSE of the competing models and proves the advantage of the overdifferenced predictor. Section 6 studies the AR(1) case using the random walk as alternative model. A simulation study is presented in section 7 supporting and illustrating the theoretical results.

2 The model and notation

Let $\{y_t\}$ be a real-valued, discrete time, series following a stationary AR$(p+1)$ process

$$\varphi(B)y_t = \alpha + a_t, \tag{2.1}$$

where B is the backshift operator; $\varphi(B) = (1 - \sum_{i=1}^{p+1} \varphi_i B^i)$ is a polynomial operator such that $\varphi(B) = 0$ has all its roots outside the unit circle; and a_t is a sequence of independent identically distributed (iid) random variables with zero mean and variance σ^2. We make the following assumption,

A1. For some $s_0 > 2$, $E\{|a_t|^{s_0}\} < \infty$.

Let denote as ρ the largest root of $\varphi(B) = 0$. We assume that the autoregressive polynomial can be factorize as $\varphi(B) = \phi(B)(1-\rho B)$, where $\phi(B) = 1-\sum_{j=1}^{p} \phi_j B^j$ and $\varphi_i = \phi_i - \rho\phi_{i-1}$, with $\phi_0 = -1$ and $\phi_{p+1} = 0$. It is well known that this model can be represented in first-order vector autoregressive form as follows

$$Y_t = A_\alpha Y_{t-1} + U_{t,p+2}, \tag{2.2}$$

with $Y_t = (y_t, ..., y_{t-p}, 1)'$, $U_{t,p+2} = (a_t, 0, ..., 0)'$, where the subindex $(p+2)$ indicates the dimension of the vectors and

$$A_\alpha = \begin{pmatrix} \varphi_1 & \varphi_2 & \cdots & \varphi_p & \varphi_{p+1} & \alpha \\ 1 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 & 1 \end{pmatrix}.$$

Then $y_t = e'_{p+2} Y_t$ with $e_{p+2} = (1, 0, ..., 0)'$. Let denote $\Gamma_y = E(Y_t Y_t')$ and $\gamma_y = E(Y_t y_{t+1})$. If we represent the process in deviations from the mean we obtain

$$\tilde{Y}_t = A_o \tilde{Y}_{t-1} + U_{t,p+1}, \tag{2.3}$$

where $\tilde{Y}_t = (\tilde{y}_t, \tilde{y}_{t-1}, ..., \tilde{y}_{t-p})'$, $\tilde{y}_t = y_t - \mu$; $\mu = E(y_t) = \varphi(1)\alpha$; and A_o is the first $(p+1) \times (p+1)$ submatrix of A_α. We will also denote as $\Gamma_{\tilde{y}} = E(\tilde{Y}_t \tilde{Y}_t')$. If a difference is

applied to y_t, the series obtained, $w_t = (1 - B)y_t$, can be represented as

$$\phi(B)(1 - \rho B)w_t = (1 - B)a_t, \tag{2.4}$$

which is noninvertible. The process w_t has the following representation (Lütkepohl 1991, page 223)

$$Z_t = A_1 Z_{t-1} + U^*_{t,p+1}, \tag{2.5}$$

with $Z_t = (W'_t, a_t)'$, $W_t = (w_t, ..., w_{t-p})'$, $U^*_{t,p+1} = (a_t, 0, \ldots, 0, a_t)'$, and

$$A_1 = \begin{pmatrix} A_o & -e_{p+1} \\ 0 \cdots 0 & 0 \end{pmatrix}$$

with $w_t = e'_{p+1} Z_t$. Let $\Gamma_w = E(W_t W'_t)$ and $\gamma_w = E(W_t w_{t+1})$. In what follows we will use the hat symbol ($\hat{o}$) to denote estimations from a sample of the overdifferenced process $\{w_t\}$ and the check symbol ($\check{o}$) for estimations from the original process $\{y_t\}$. The least square estimator for the AR($p+1$) parameters $\varphi = (\varphi_1, ..., \varphi_{p+1}, \alpha)'$ fitted from a sample of size T of the original process (2.1) is

$$\check{\varphi} = \check{\Gamma}_y^{-1} \check{\gamma}_y, \tag{2.6}$$

where $\check{\Gamma}_y = (T - p - 1)^{-1} \sum_{j=p+1}^{T-1} Y_j Y'_j$, $\check{\gamma}_y = (T - p - 1)^{-1} \sum_{j=p+1}^{T-1} Y_j y_{j+1}$. Similarly the least square estimator of the parameters $\phi = (\phi_1, ..., \phi_p)'$ from a misspecified AR(p) fitted from a sample of size $T - 1$ ($t = 2, 3, ..., T$) of the overdifferenced process (2.4) is

$$\hat{\phi} = \hat{\Gamma}_w^{-1} \hat{\gamma}_w, \tag{2.7}$$

where $\hat{\Gamma}_w = (T - p - 1)^{-1} \sum_{j=p+1}^{T-1} W_j W'_j$, $\hat{\gamma}_w = (T - p - 1)^{-1} \sum_{j=p+1}^{T-1} W_j w_{j+1}$. We make, further, the following assumptions:

A2. $E\{\|\check{\Gamma}_y^{-1}\|^{2k}\}$ ($k = 1, 2, ..., k_0$) is bounded for $T > T_0$ and some k_0.
A3. $E\{\|\hat{\Gamma}_w^{-1}\|^{2k}\}$ ($k = 1, 2, ..., k_0$) is bounded for $T > T_0$ and some k_0.

Assumptions A2 and A3 are similar to assumtion A3 by Kunitomo & Yamamoto (1985) and are satisfied if a_t is normal.

3 Overdifferencing a nearly nonstationary autoregression

3.1 General considerations

In this section we will analyze the properties of the estimator $\hat{\phi} = \hat{\Gamma}_w^{-1}\hat{\gamma}_w$ for the misspecified ARIMA$(p,1,0)$ when the process is nearly nonstationary. In general, a time series is said to be nearly nonstationary (near integrated) if its largest root, ρ, is very close to unity. This idea has been formalize in the statistical literature (Phillips 1987) by reparameterizing this largest root as

$$\rho = \exp\left(-\frac{c}{T}\right) = 1 - \frac{c}{T} + o(T^{-1}), \tag{3.1}$$

where c is a fixed constant and T is the sample size. The limitation of definition (3.1), for our purpose, is that the convergence rate to unity is fixed to be $O(T^{-1})$. The reason of this rate is that it is the order of consistency of the least square estimator of a unit root. In this paper we will employ a more general definition by writting ρ, the largest root of the process (2.1), as

$$\rho = \exp\left(-\frac{c}{T^{\beta}}\right), \tag{3.2}$$

with c, β being fixed constants. We deal only with the case $c > 0$, where the largest root is lower than unity but approach this value at a convergence rate $O(T^{-\beta})$.

Let $\pi(B)w_t = a_t$ be the autoregressive form of the overdifferenced process (2.4). The coefficients of $\pi(B)$ follow

$$\pi_j = \begin{cases} \phi_j + (\rho - 1)(1 - \sum_{k=1}^{j-1} \phi_k) & \text{if } j \leq p \\ (\rho - 1)(1 - \sum_{k=1}^{p} \phi_k) & \text{if } j > p. \end{cases} \tag{3.3}$$

If ρ follows (3.2) with β large enough, the term $(\rho - 1)$ will be small ($O(T^{-\beta})$) compared to the sampling variability of estimated correlograms (the standard error of sampling autocorrelation coefficients is $O(T^{-\frac{1}{2}})$). Therefore, although the overdifferenced process w_t is strictly a noninvertible ARMA$(p+1,1)$, an average correlogram of w_t will suggest to estimate an AR(p) instead. Figure 1 shows the result of a simulation study to illustrate this point. In each replication of the simulation we have calculated the estimated autocorrelation function (*acf*) and partial autocorrelation function (*pacf*) of both

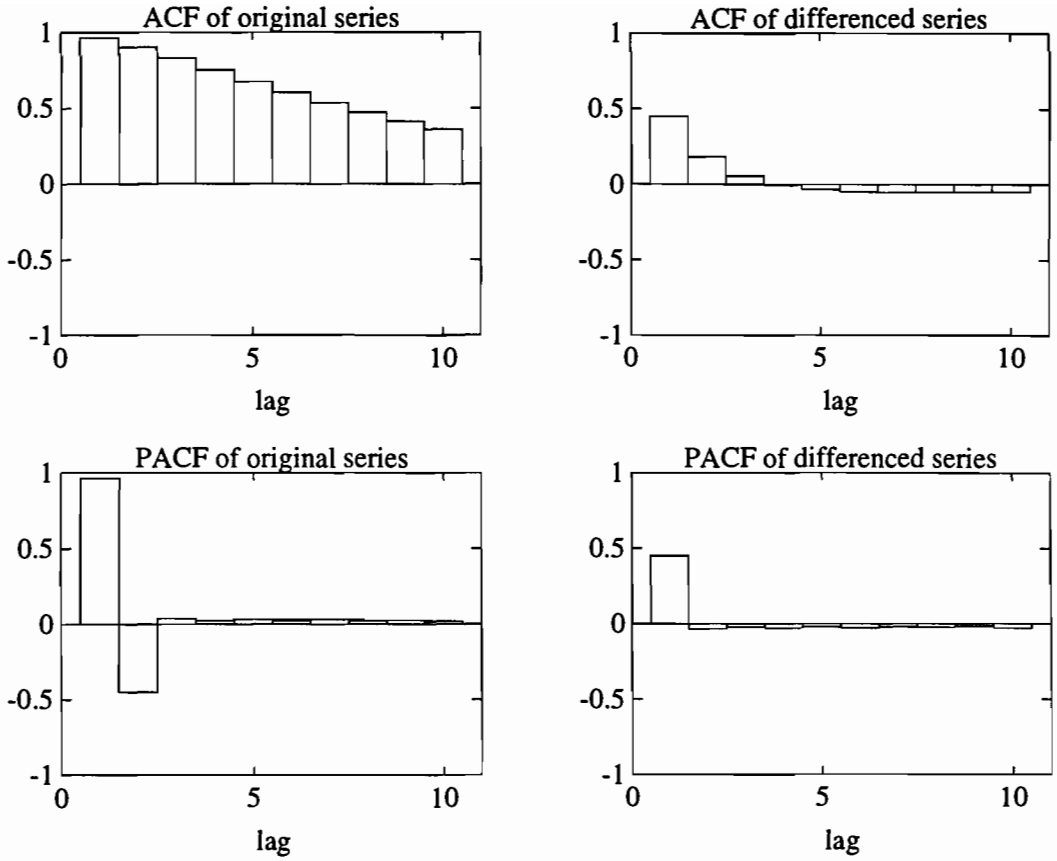

Figure 1: Estimated *acf* and *pacf* of the model $(1-0.5B)(1-0.95B)y_t = 10+a_t$, a_t iid and following a $N(0,1)$, sample size $T = 100$. Average of 5000 replications.

the original series and the differenced series of lenght $T = 100$. The simulated model is $(1 - 0.5B)(1 - 0.95B)y_t = 10 + a_t$ where a_t is an iid process following $a_t \sim N(0,1)$. The graph is the result of averaging the correlograms of 5000 replications. This figure shows that the more plausible modellizations are an AR(2) and an ARIMA$(1,1,0)$. This approach of fitting an AR(p) instead of an ARMA$(p+1,1)$ is equivalent to estimate a truncation of order p of an infinite order autoregresion with coefficients (3.3). Berk (1974) and Banshali (1978) analyze the truncation of a possibly infinite order autoregression when the process is both stationary and invertible and they find the order of the truncation that allows to ignore the bias of the misspecification. In this paper we deal with a noninvertible process and a truncation made at a fixed place (order p). We investigate, then, the properties of the process in order to obtain both consistent estimates of the proposed model and efficient predictors, and therefore ignore also the bias of the misspecification.

The expression (3.3) also reveals the influence of the remaining roots in small samples. If we denote as r_i, $i = 1, ..., p$ to the roots of the characteristic function $\phi(B) = 0$ then $\phi(B) = \prod_{i=1}^{p}(1 - r_i B)$. For $B = 1$ it can be written

$$(1 - \sum_{k=1}^{p} \phi_k) = \prod_{i=1}^{p}(1 - r_i). \tag{3.4}$$

Therefore although the departure of π_j from ϕ_j depends mainly on $(1 - \rho)$ it is influenced by the remaining roots. Negative values of r_i increase the value of π_j, $j > p$ and increase the bias of the proposed truncation at $j = p$ in small sample sizes.

3.2 Root-T consistency of $\hat{\phi}$.

Let us denote as $\{w_{t|p}\}$ to the limit process of $\{w_t\}$ when $T \to \infty$ and therefore $\rho \to 1$. This limit process follows a pure AR(p) process with markovian representation

$$W_{t|p} = A_p W_{t-1|p} + U_{t,p}, \tag{3.5}$$

where A_p is a $p \times p$ matrix with the same structure than A_o but with the coefficients $(\phi_1, \ldots, \phi_p)$ in the first row;$W_{t|p} = (w_{t|p}, ..., w_{t-p+1|p})'$. Then we have from (2.4)

$$\begin{aligned} w_t &= \phi^{-1}(B)(1 - \rho B)^{-1}(1 - B)a_t \\ &= \phi^{-1}(B)\left[1 - (1 - \rho)(B + \rho B^2 + \cdots)\right] a_t \\ &= w_{t|p} - \sum_{j=0}^{\infty} \psi_j (1 - \rho) z_{t-1-j}, \end{aligned} \tag{3.6}$$

where ψ_j ; $j = 0, 1, \ldots$ are the coefficients of $\phi^{-1}(B)$, and $(1 - \rho B)z_t = a_t$. Let us denote as $\Gamma_{w|p} = E(W_{t|p} W'_{t|p})$ and $\gamma_{w|p} = E(W_{t|p} w_{t+1|p})$. We define also the sampling autocovariances as $\hat{\Gamma}_{w|p} = (T - p - 1)^{-1} \sum_{j=p+1}^{T-1} W_{j|p} W'_{j|p}$, $\hat{\gamma}_{w|p} = (T - p - 1)^{-1} \sum_{j=p+1}^{T-1} W_{j|p} w_{j+1|p}$, and also make the following assumption:

A3'.$E\{\|\hat{\Gamma}_{w|p}^{-1}\|^{2k}\}$ $(k = 1, 2, ..., k_0)$ is bounded for $T > T_0$ and some k_0.

Since the elements of both $\hat{\Gamma}_w$ and $\hat{\gamma}_w$ are sampling autocovariances we obtain that

$$\begin{aligned} \hat{\Gamma}_w &= \hat{\Gamma}_{w|p} + O_p(r_t), \\ \hat{\gamma}_w &= \hat{\gamma}_{w|p} + O_p(r_t). \end{aligned}$$

where $r_t = \sum_{j=0}^{\infty} \psi_j (1-\rho) z_{t-1-j}$. The magnitude of the error term r_t is determined in the following theorem.

Theorem 1 *Let $\{w_t\}$ be a time series generated by (2.4) and let $w_1, ..., w_T$ be a sample from this process. Let its largest root ρ follows*

$$\rho = exp\left(-\frac{c}{T^{\beta}}\right) \quad ; \ \beta > 1. \tag{3.7}$$

Then

$$w_t = w_{t|p} + o_p(T^{-\frac{1}{2}}) \tag{3.8}$$

and

$$\hat{\Gamma}_w = \hat{\Gamma}_{w|p} + o_p(T^{-\frac{1}{2}}), \tag{3.9}$$

$$\hat{\gamma}_w = \hat{\gamma}_{w|p} + o_p(T^{-\frac{1}{2}}). \tag{3.10}$$

Also, if $\beta = 1$, the probability order in (3.8), (3.9) and (3.10) is $O_p(T^{-\frac{1}{2}})$.

Proof: Since

$$E(z_t^2) = \frac{\sigma^2}{1-\rho^2} = O\left((1-\rho^2)^{-1}\right), \tag{3.11}$$

then, by Chebyshev's inequality $z_t = O_p\left((1-\rho^2)^{-\frac{1}{2}}\right)$. Thus, since w_t is stationary,

$$O_p(r_t) = O_p\left(\left[\frac{1-\rho}{1+\rho}\right]^{\frac{1}{2}}\right). \tag{3.12}$$

Applying that $\rho = \exp(-c/T^{\beta}) = 1 - c/T^{\beta} + O(T^{-2\beta})$, we obtain

$$\frac{1-\rho}{1+\rho} = O(T^{-\beta}). \tag{3.13}$$

Since $\beta \geq 1$ the theorem holds. □

Although we have imposed the definition of ρ in (3.2) it is easily verified that it appears in a natural fashion in this context. Let us denote $T^{-\beta} = (1-\rho)/(1+\rho)$. Then

$$\rho = \frac{T^{\beta}-1}{T^{\beta}+1} = 1 - \frac{2}{T^{\beta}+1}.$$

Since $T^{-\beta} < 1$

$$\rho = 1 - \frac{2}{T^{\beta}}\left[\sum_{j=0}^{\infty}\left(\frac{-1}{T^{\beta}}\right)^j\right] = e^{-\frac{2}{T^{\beta}}} + o(T^{-3\beta}).$$

The term $O\left((1-\rho)(1+\rho)^{-1}\right)$ in (3.12) is not affected by the constant term of the exponential and the number 2 has been replaced by the constant c in the definition of ρ

Corollary 1 *Let conditions of theorem 1 hold, with* $\beta \geq 1$ *then*

$$\begin{aligned} \hat{\gamma}_w &= \gamma_{w|p} + O_p(T^{-\frac{1}{2}}), \\ \hat{\Gamma}_w &= \Gamma_{w|p} + O_p(T^{-\frac{1}{2}}). \end{aligned}$$

Proof: Given that $w_{t|p}$ is a stationary process it holds that $\hat{\gamma}_{w|p} = \gamma_{w|p} + O_p(T^{-\frac{1}{2}})$. Applying this to theorem 1 the results hold. □

We can now prove the root-T consistency of $\hat{\phi}$.

Theorem 2 *Let the conditions of theorem 1 hold with* $\beta \geq 1$ *, then*

$$\hat{\phi} = \phi + O_p(T^{-\frac{1}{2}}).$$

See proof in appendix B.

3.3 Bias and mean squared error of $\hat{\phi}$.

Let $\hat{\phi}_{|p}$ be the least square estimator of ϕ from a sample of a true ARIMA$(p,1,0)$ process. The bias and mean square error (MSE) of this estimator from a properly specified autoregression has largely been investigated and can be found in the works of Marriot & Pope (1954), Kendall (1954), Whitte (1961), Shenton & Johnson (1965), Bhansali (1981), Hosoya & Taniguchi (1982), Tjøsteim & Paulsen (1983), Tanaka (1984), Yamamoto & Kunitomo (1984), Kunitomo & Yamamoto (1985) and Shaman & Stine (1988) among others. Since the similarity of the estimators $\hat{\phi}$ and $\hat{\phi}_{|p}$ depends on the near nonstationarity hypothesis we will express their differences in terms of ρ. The following theorems formulate the first and second moments of the least square estimator $\hat{\phi}$, of a near nonstationary overdifferenced AR$(p+1)$ process, around the true parameter ϕ as the first and second moments of $\hat{\phi}_{|p}$ around ϕ plus an error term depending on ρ.

Theorem 3 *Assume A1, A2 (with* $s_o = 16$*), A3 and A3'. Then*

$$E(\hat{\phi} - \phi) = E(\hat{\phi}_{|p} - \phi) + O\left(\left[\frac{1-\rho}{1+\rho}\right]^{\frac{1}{2}}\right). \tag{3.14}$$

The proof is given in appendix B. Since $((1-\rho)(1+\rho)^{-1}) = O(T^{-\beta})$ and given that $E(\hat{\phi}_{|p} - \phi) = O(T^{-1})$ (see, for instance, Bhansali 1981) we need a value $\beta > 2$ if we want that the biases are equal up to terms $O(T^{-1})$, whereas for root-T consistency we only need $\beta \geq 1$.

Theorem 4 *Assume A1 (with $s_o = 16$), A2, A3 and A3'. Then*

$$E[(\hat{\phi}-\phi)(\hat{\phi}-\phi)'] = E[(\hat{\phi}_{|p}-\phi)(\hat{\phi}_{|p}-\phi)'] + O\left(max\left\{\left(\frac{1-\rho}{1+\rho}\right)^{\frac{1}{2}} T^{-\frac{1}{2}}, \frac{1-\rho}{1+\rho}\right\}\right)$$

See proof in appendix B. We can see from this theorem that the MSE of both predictors are closer to each other than the biases. If ρ is such that $\beta > 1$ then both expresions of MSE are equal up to $O(T^{-1})$.

4 Mean squared error of H-step prediction

In this section we will obtain the expresions of the mean square error of predicting y_{T+H} form $t = T$. In order to compare the PMSE of the AR$(p+1)$ model with the PMSE of the overdifferenced AR$(p,1,0)$ model we need to reexpress their estimated H-steps ahead predictions ($\check{y}_{T+H}$ and $\hat{y}_{T+H}$) in terms of the estimated increments ($\check{w}_t$ and $\hat{w}_t$ respectively). For the AR$(p+1)$ model the estimated increments are $\check{w}_t = \check{y}_t - \check{y}_{t-1}$. Then

$$\check{y}_{T+H} = y_T + \sum_{h=1}^{H} \check{w}_{T+h}. \tag{4.1}$$

Hence

$$\begin{aligned} \text{PMSE}(\check{y}_{T+H}) &= E(y_{T+H} - \check{y}_{T+H})^2 \\ &= \sum_{h=1}^{H} \text{PMSE}(\check{w}_{T+h}) + 2\sum_{h=1}^{H}\sum_{k=h+1}^{H} \text{E}\left[(w_{T+h} - \check{w}_{T+h})(w_{T+k} - \check{w}_{T+k})\right]. \end{aligned} \tag{4.2}$$

In the same way, for the overdifferenced model, the PMSE$(\hat{y}_{T+H})$ can be expressed as

$$\text{PMSE}(\hat{y}_{T+H}) = \sum_{h=1}^{H} \text{PMSE}(\hat{w}_{T+h}) + 2\sum_{h=1}^{H}\sum_{k=h+1}^{H} \text{E}\left[(w_{T+h} - \hat{w}_{T+h})(w_{T+k} - \hat{w}_{T+k})\right]. \tag{4.3}$$

4.1 PMSE of the properly specified AR(p+1) predictor

Let denote as $\check{A}_\alpha$ the least square estimation of A_α in the sample y_1, y_2,..., y_T, using the properly specified model (2.2). The predicted value $\check{y}_{T+H}$ using this information is

$$\check{y}_{T+H} = e'_{p+2}\check{A}_\alpha^H Y_T. \tag{4.4}$$

According with Kunitomo & Yamamoto (1985), the PMSE of predicting y_{T+H} from T, using this unbiased estimated predictor is

$$\begin{aligned}\text{PMSE}(\check{y}_{T+H}) &= \sigma^2 \sum_{h=0}^{H-1}(e'_{p+2}A_\alpha^h e_{p+2})^2 + \frac{\sigma^2}{T}\sum_{h=0}^{H-1}\sum_{k=0}^{H-1}(e'_{p+2}A_\alpha^h e_{p+2})(e'_{p+2}A_\alpha^k e_{p+2}) \\ &\quad \times \text{trace}\left(A_\alpha^{H-1-h}\Gamma_y A_\alpha'^{H-1-k}\Gamma_y^{-1}\right) + O(T^{-3/2}).\end{aligned} \tag{4.5}$$

Nevertheless in order to compare the predictive performance of this modellization with the overdifferenced model ARIMA$(p,1,0)$ we need to express the PMSE in terms of the overdifferenced series w_t as shown in (4.2). The estimated increment $\check{w}_{T+h}$ as a function of the estimated coefficients $\check{A}_\alpha$ are

$$\check{w}_{T+h} = \check{y}_{T+h} - \check{y}_{T+h-1} = e'_{p+2}\check{A}_\alpha^{h-1}(\check{A}_\alpha - I_{p+2})Y_T \tag{4.6}$$

where I_{p+2} is the $(p+2)\times(p+2)$ identity matrix. The observed value w_{T+h} is

$$w_{T+h} = e'_{p+2}A_\alpha^{h-1}(A_\alpha - I_{p+2})Y_T + L_1 - L_2,$$

where

$$\begin{aligned}L_1 &= \sum_{k=0}^{h-1} e'_{p+2}A_\alpha^k U_{t+h-k,p+2}, \\ L_2 &= \sum_{k=0}^{h-2} e'_{p+2}A_\alpha^k U_{t+h-1-k,p+2} = \sum_{k=1}^{h-1} e'_{p+2}A_\alpha^{k-1}U_{t+h-k,p+2}.\end{aligned}$$

The prediction error is then

$$\check{w}_{T+h} - w_{T+h} = e'_{p+2}(\check{A}_\alpha^h - A_\alpha^h)Y_T - e'_{p+2}(\check{A}_\alpha^{h-1} - A_\alpha^{h-1})Y_T - L_1 + L_2 \tag{4.7}$$

The PMSE$(\check{w}_{T+h})$ and $\text{E}[(\check{w}_{T+h} - w_{T+h})(\check{w}_{T+k} - w_{T+k})]$ are shown in the following theorem (see proof in appendix C).

Theorem 5 *Let w_t follows (2.4) with parameter $\rho = exp(-c/T^{\beta})$, with $\beta > 1$. Let assume A2, A3, A3' and A1 with $s_0 = 32$ when $h = 1, 2$ and $s_0 = 16h$ when $h \geq 3$. Then the h steps ahead predictive mean squared error using the estimated predictor (4.6) is*

$$PMSE(\check{w}_{T+h}) = E(\check{w}_{T+h} - w_{T+h})^2 = \sigma^2 \sum_{j=0}^{h-1} (e'_{p+2} A_1^j c_{p+2})^2 + \frac{\sigma^2}{T} \sum_{j=0}^{h-1} \sum_{k=0}^{h-1} (e'_p A_p^j e_p)(e'_p A_p^k e_p) \times trace\left(A_\alpha^{h-1-j} \Gamma_y A_\alpha'^{h-1-k} \Gamma_y^{-1}\right) + o(T^{-1}), \quad (4.8)$$

and, for $k \geq h$,

$$E[(\check{w}_{T+h} - w_{T+h})(\check{w}_{T+k} - w_{T+k})] = \sigma^2 \sum_{i=0}^{h-1} (e'_{p+2} A_1^i c_{p+2})(e'_{p+2} A_1^{i+(k-h)} c_{p+2}) + \frac{\sigma^2}{T} \sum_{n=0}^{k-1} \sum_{i=0}^{h-1} (e'_p A_p^n e_p)(e'_p A_p^i e_p) \times trace\left(A_\alpha^{h-1-i} \Gamma_y A_\alpha^{k-1-n} \Gamma_y^{-1}\right) + o(T^{-1}), \quad (4.9)$$

where $c_{p+1} = (1, 0, ..., 0, 1)'$.

The terms at the right side of (4.8) and (4.9) have two components. The first part is the variance of the prediction errors and the covariance between prediction errors at different horizonts, respectively, of the true ARIMA$(p+1, 1, 1)$ process. The second part is the sampling error due to the estimation of the $p + 2$ parameters $\hat{\varphi}$. The terms $(e'_p A_p^v e_p); v = j, k, n, i,$ are approximations of the terms $(e'_{p+2} A_1^v e_{p+2}); v = j, k, n, i$. This sustitution, under the assumption of theorem 5, causes an error of low magnitude order, $o(T^{-1})$, but allows us the comparison with the PMSE of the overdifferenced model.

4.2 PMSE of the overdifferenced ARIMA(p,1,0) predictor.

Let us assume that we employ as predictor for w_{T+h} the one derived from the estimated ARIMA$(p, 1, 0)$, that is

$$\hat{w}_{T+h} = e'_p \hat{A}_p^h W_T \quad (4.10)$$

where $\hat{A}_p$ is the least square estimator of A_p.

We can rewrite (4.10) as

$$\begin{aligned} \hat{w}_{T+h} &= e'_p A_p^h W_T + e'_p (\hat{A}_p^h - A_p^h) W_T \\ &= E(w_{T+h|p}|T) + e'_p (\hat{A}_p^h - A_p^h) W_T \end{aligned}$$

The true value w_{T+h} is, from (2.5)

$$w_{T+h} = e'_{p+2}A_1^h Z_T + L_h = E(w_{T+h}|T) + L_h$$

where $L_h = \sum_{j=0}^{h-1} e'_{p+2}A^j U^*_{T+h-j}$. By (3.6)

$$E(w_{T+h}|T) = E(w_{T+h|p}|T) - \sum_{j=h-1}^{\infty} \psi_j(1-\rho)z_{T-1-j} - \sum_{j=0}^{h-2} \psi_j(1-\rho)\rho^{h-1-j}z_T \qquad (4.11)$$

Then, the h steps ahead prediction error of the predictor (4.12) is then

$$\begin{aligned}(w_{T+h} - \hat{w}_{T+h}) &= L_h - e'_p(\hat{A}_p^h - A_p^h)W_T \\ &\quad - \sum_{j=h-1}^{\infty} \psi_j(1-\rho)z_{T-1-j} - \sum_{j=0}^{h-2} \psi_j(1-\rho)\rho^{h-1-j}z_T \qquad (4.12)\end{aligned}$$

The following theorem gives an approximation of order $o(T^{-1})$ of the expectation of the lead-h predictive square error (see proof in appendix C).

Theorem 6 *Let w_t follows (2.4) with parameter $\rho = exp(-c/T^\beta)$, with $\beta > 1$. Let assume A2, A3, A3' and A1 with $s_0 = 32$ when $h = 1,2$ and $s_0 = 16h$ when $h \geq 3$. Then the h steps ahead predictive mean squared error using the predictor (4.10) is*

$$\begin{aligned}PMSE(\hat{w}_{T+h}) = E(\hat{w}_{T+h} - w_{T+h})^2 &= \sigma^2 \sum_{k=0}^{h-1}(e'_{p+2}A_1^k c_{p+2})^2 + \frac{\sigma^2}{T}\sum_{j=0}^{h-1}\sum_{k=0}^{h-1}(e'_p A_p^j e_p)(e'_p A_p^k e_p) \\ &\quad \times trace\left(A_p^{h-1-j}\Gamma_{w|p}A_p'^{h-1-k}\Gamma_{w|p}^{-1}\right) + o(T^{-1}), \qquad (4.13)\end{aligned}$$

and, for $k \geq h$,

$$\begin{aligned}E[(\hat{w}_{T+h} - w_{T+h})(\hat{w}_{T+k} - w_{T+k})] &= \sigma^2 \sum_{i=0}^{h-1}(e'_{p+2}A_1^i c_{p+2})(e'_{p+2}A_1^{i+(k-h)}c_{p+2}) \\ &+ \frac{\sigma^2}{T}\sum_{n=0}^{k-1}\sum_{i=0}^{h-1}(e'_p A_p^n e_p)(e'_p A_p^i e_p) \times trace\left(A_p^{h-1-i}\Gamma_{w|p}A_p^{k-1-n}\Gamma_{w|p}p^{-1}\right) + o(T^{-1}), \qquad (4.14)\end{aligned}$$

where $c_{p+2} = (1,0,...,0,1)'$.

As mentioned in the previous subsection, the terms at the right side of (4.13) and (4.14) have two components. The first part, the variance of prediction errors and their covariance between different horizonts of the true ARIMA$(p+1,1,1)$ process, is the same than in theorem 5. The second part is the sampling error due to the estimation of the p parameters $\hat{\phi}$, in contrast with the estimation of the $p+2$ parameters of the AR$(p+1)$ model. It should be observed that this second components differ from the ones on the previous subsection only in the elements inside the trace operators.

5 Comparing the prediction accuracy

In this section we compare the PMSE found in the last section for the two models. We prove that, under the assumption of near nonstationarity exposed in (3.7), overdifferencing produces lower PMSE. Comparing expressions on theorem 5 and theorem 6 it can be seen that the only difference between PMSE($\check{y}_{T+H}$) and PMSE($\hat{y}_{T+H}$) is in the elements inside the trace operators. These differences are solved applying the following lemmas: lemma 1 compares such a trace in processes with and without constant term; lemma 2 compares the trace in nearly nonstationary processes with no constant term and the overdifferenced one. The proofs of these lemmas can be found in appendix D.

Lemma 1 *Let y_t follows the process (2.1) with $\alpha \neq 0$. Then*

$$trace\left(A_\alpha^i \Gamma_y A_\alpha'^j \Gamma_y^{-1}\right) = 1 + trace\left(A_o^i \Gamma_{\check{y}} A_o'^j \Gamma_{\check{y}}^{-1}\right). \tag{5.1}$$

Lemma 2 *Let y_t follows the process (2.1) with largest root $\rho = exp(-c/T^\beta)$; $\beta > 1$. Then*

$$trace(A_o^i \Gamma_{\check{y}} A_o'^j \Gamma_{\check{y}}^{-1}) = \rho^{i+j} + trace\left(A_p^i \Gamma_{w|p} A_p^j \Gamma_{w|p}^{-1}\right) + o(T^{-1}). \tag{5.2}$$

It is posible now to prove the advantages of overdifferencing when the process is nearly nonstationary.

Theorem 7 *Let y_t follows the process (2.1) and let the conditions of theorems 5 and 6 hold. Then, for $H \geq 1$,*

$$PMSE(\hat{y}_{T+H}) < PMSE(\check{y}_{T+H}). \tag{5.3}$$

Proof: By direct application of lemma 1 and lemma 2 to the differences of (4.8) with (4.13) and expression (4.14) with (4.9) we obtain (see (4.2) and (4.3))

$$\text{PMSE}(\check{y}_{T+H}) - \text{PMSE}(\hat{y}_{T+H}) = \frac{\sigma^2}{T}\left(\sum_{h=1}^{H}\sum_{j=0}^{h-1}\psi_j\right)^2 + \frac{\sigma^2}{T}\left(\sum_{h=1}^{H}\sum_{j=0}^{h-1}\psi_j \rho^{h-1-j}\right)^2 > 0, \tag{5.4}$$

where $\psi_j = (e_p' A_p^j e_p)$, $j = 1, ..., H$. □

Expression (5.4) shows that the advantages of the overdifferenced model can be decomposed into two parts. The first term at the right side of (5.4) is the result of applying

lemma 1 and hence is the difference due to the estimation of the constant term α in the AR($p+1$) model. The second term is the result of applying lemma 2 and therefore is due to the estimation of an extra parameter in the autoregressive polynomial. Then the superior forecasting performance of the model ARIMA($p,1,0$) is due to its more parsimonious representation. The difference (5.4) increases monotonically with the predicting horizon. For $H = 1$ the difference is $2\sigma^2/T$ if a constant $\alpha \neq 0$ is needed, and σ^2/T if $\alpha = 0$ and no constant is estimated. This result is similar to Ledolter & Abraham (1981), where they state that each innecesary estimated parameter increases the one step ahead PMSE by σ^2/T. In our context this is equivalent to say that both α and ρ are innecessary parameters and therefore there is a loss in efficiency (that tends to zero asymptotically) in the full parameterizated model AR($p+1$). For $H > 1$ the loss in efficiency increases but in an amount that depends on the model through the coefficients ψ_j.

Conversely, if $\beta < 1$ the overdifferenced model has a worse forecasting performance than the true model as stated in the following theorem.

Theorem 8 *Let y_t follows the process (2.1) and let the conditions of theorem 1 hold with $\beta < 1$. Then, for $h > 0$,*

$$PMSE(\hat{y}_{T+H}) > PMSE(\check{y}_{T+H}). \tag{5.5}$$

See proof in appendix D.

6 A simpler case: overdifferencing AR(1) processes

Although results in previous sections are aplicable to a general stationary autoregresion it is usefull to analize the AR(1) process. One reason is that the formulation of PMSE is simpler and some asymptotic approximations used in last sections are not necessary. Besides, results will not be afected by any other root, as shown in (3.4), and can be considered as a neutral benchmarck. We will analyze both the AR(1) case with no intercept (AR(1)) and with intercept (AR(1,μ)).

6.1 The AR(1) case.

Let y_t, $t = 1, 2, ..., T$, be a sample of the stationary process

$$y_t = \phi y_{t-1} + a_t \ , \phi < 1, \tag{6.1}$$

and let denote as $\check{\phi}$ the least square estimator of ϕ. The conditional expectation of y_{T+H}, given information until time T is $\check{y}_{T+H} = \check{\phi}^H y_T$, and the PMSE is (see (4.4))

$$\text{PMSE}(y_{T+H}|\text{AR}(1)) = \sigma^2 \left[\frac{1-\phi^{2H}}{1-\phi^2} + \frac{H^2 \phi^{2H-2}}{T} \right] + O(T^{-\frac{3}{2}}). \tag{6.2}$$

The overdifferenced predictor is the random walk (RW). Then, the prediction of y_{T+H} from T is $\hat{y}_{T+H} = y_T$ and therefore

$$\text{PMSE}(y_{T+H}|\text{RW}) = \sigma^2 \left(\frac{1-\phi^{2H}}{1-\phi^2} + \frac{(1-\phi^H)^2}{1-\phi^2} \right). \tag{6.3}$$

Comparing (6.2) and (6.3) it can be verified that, if the inequality

$$\frac{H^2 \phi^{2(H-1)}}{T} - \frac{(1-\phi^H)^2}{1-\phi^2} > 0 \tag{6.4}$$

holds, then $\text{PMSE}(y_{T+H}|\text{AR}(1)) > \text{PMSE}(y_{T+H}|\text{RW})$. Table 6.1 shows the relationship between ϕ, T and H in order to fulfill the innequality (6.4). This table shows that as the horizon increases it is more difficult to overcome the correct AR(1) model, nevertheless this variation with H is very small. To study this effect we can see for $H = 1$ that, by (6.4)

$$\frac{1}{T} > \frac{1-\phi}{1+\phi}, \tag{6.5}$$

and therefore

$$\phi > 1 - \frac{2}{T+1} = \exp\left(-\frac{2}{T}\right) + O(T^{-3}). \tag{6.6}$$

Since the convergence rate to unity is $O(T^{-1})$ and using a Taylor expansion we obtain

$$\phi^H = 1 - H(1-\phi) + O(T^{-2}). \tag{6.7}$$

Let denote as $\phi(H)$ to the minimum coefficient that fulfills the inequality (6.4) at horizon H. Then

$$\phi(H) = 1 - \frac{2}{T+1+4(H-1)} + O(T^{-2}) = \exp\left(-\frac{2}{T+4(H-1)}\right) + O(T^{-2}). \tag{6.8}$$

Table 1: Lowest values of ϕ to obtain PMSE(y_{T+H}|AR(1)) > PMSE(y_{T+H}|RW)

	Horizon H							
T	1	2	3	4	5	10	15	20
25	0.923	0.929	0.933	0.937	0.940	0.951	0.958	0.963
50	0.961	0.962	0.964	0.965	0.966	0.970	0.973	0.976
75	0.974	0.974	0.975	0.976	0.976	0.978	0.980	0.982
100	0.980	0.981	0.981	0.981	0.982	0.983	0.984	0.985
150	0.987	0.987	0.987	0.987	0.987	0.988	0.989	0.989
200	0.990	0.990	0.990	0.990	0.990	0.991	0.991	0.992
300	0.993	0.993	0.993	0.994	0.994	0.994	0.994	0.994
500	0.996	0.996	0.996	0.996	0.996	0.996	0.996	0.996

For instance, if $\phi = 0.97$ and $T = 50$ the random walk will have lower PMSE than the AR(1) until $H = 10$. From this horizon onward the RW will have a slightly larger PMSE than the correct predictor. It is easily verified that $\phi(H) - \phi(1) = O(HT^{-2})$ and $\phi(H+1) - \phi(H) = O(T^{-2})$. Therefore the dependency of $\phi(H)$ with the forecasting horizon is of magnitude $O(T^{-2})$ and only appreciable in small sample sizes and large horizons.

6.2 The AR(1,μ) case.

Let y_t, $t = 1, 2, ..., T$, be a sample of the stationary process

$$y_t = \alpha + \phi y_{t-1} + a_t \quad , \phi < 1, \tag{6.9}$$

and let denote as $\check{\alpha}$ and $\check{\phi}$ the least square estimators of α and ϕ respectively. Following the same arguments than in the preceding subsection we have, using information until T

$$E(y_{T+H}|T) = \check{y}_{T+H} = \check{\alpha}\frac{1-\check{\phi}^H}{1-\check{\phi}} + \check{\phi}^H y_T, \tag{6.10}$$

and the PMSE of this predictor, by (4.4), is

$$\text{PMSE}(y_{T+H}|\text{AR}(1|\mu)) = \sigma^2\frac{1-\phi^{2H}}{1-\phi^2} + \frac{\sigma^2}{T}\left[H^2\phi^{2H-2} + \left(\frac{1-\phi^H}{1-\phi}\right)^2\right] + O(T^{-\frac{3}{2}}). \tag{6.11}$$

Table 2: Lowest values of ϕ to obtain PMSE(y_{T+H}|AR(1 | μ)) > PMSE(y_{T+H}|RW)

	Horizon H							
T	1	2	3	4	5	10	15	20
25	0.852	0.862	0.869	0.876	0.881	0.898	0.907	0.913
50	0.923	0.926	0.928	0.931	0.932	0.940	0.945	0.948
75	0.948	0.949	0.951	0.952	0.953	0.957	0.960	0.962
100	0.961	0.962	0.962	0.963	0.964	0.966	0.968	0.970
150	0.974	0.974	0.974	0.975	0.975	0.976	0.977	0.978
200	0.980	0.980	0.981	0.981	0.981	0.982	0.982	0.983
300	0.987	0.987	0.987	0.987	0.987	0.988	0.988	0.988
500	0.992	0.992	0.992	0.992	0.992	0.992	0.992	0.992

Conversely if the predictor is the RW it follows that

$$\begin{aligned} \text{PMSE}(y_{T+H}|\text{RW}) &= \sigma^2\frac{1-\phi^H}{1-\phi^2}+\frac{\alpha^2(1-\phi^H)^2}{(1-\phi)^2}+(1-\phi^H)^2E(y_T^2) \\ &\quad -2\frac{\alpha(1-\phi^H)^2}{1-\phi}E(y_T). \end{aligned} \tag{6.12}$$

Since $E(y_T)=\frac{\alpha}{1-\phi}$, and $E(y_T^2)=\frac{\sigma^2}{1-\phi^2}+\left(\frac{\alpha}{1-\phi}\right)^2$, we obtain

$$\text{PMSE}(y_{T+H}|\text{RW})=\sigma^2\left(\frac{1-\phi^{2H}}{1-\phi^2}+\frac{(1-\phi^H)^2}{1-\phi^2}\right), \tag{6.13}$$

which coincides with (6.3). The comparison of (6.11) with (6.13) provides that if

$$\frac{H^2\phi^{2(H-1)}}{T}+\frac{(1-\phi^H)^2}{T(1-\phi)^2}-\frac{(1-\phi^H)^2}{1-\phi^2}>0, \tag{6.14}$$

then PMSE(y_{T+H}|AR(1 | μ)) > PMSE(y_{T+H}|RW). Table (6.1) shows the range values of ϕ determined by this inequality. Using (6.7) we obtain for $H \geq 1$

$$\phi(H)=1-\frac{4}{T+2+4(H-1)}+O(T^{-2})=\exp\left(-\frac{4}{T+4(H-1)}\right)+O(T^{-2}). \tag{6.15}$$

We observe in table (6.1) and also in expression (6.15) that it is easier for the RW to overcome the autoregressive model than in the case with no intercept. This reinforces the

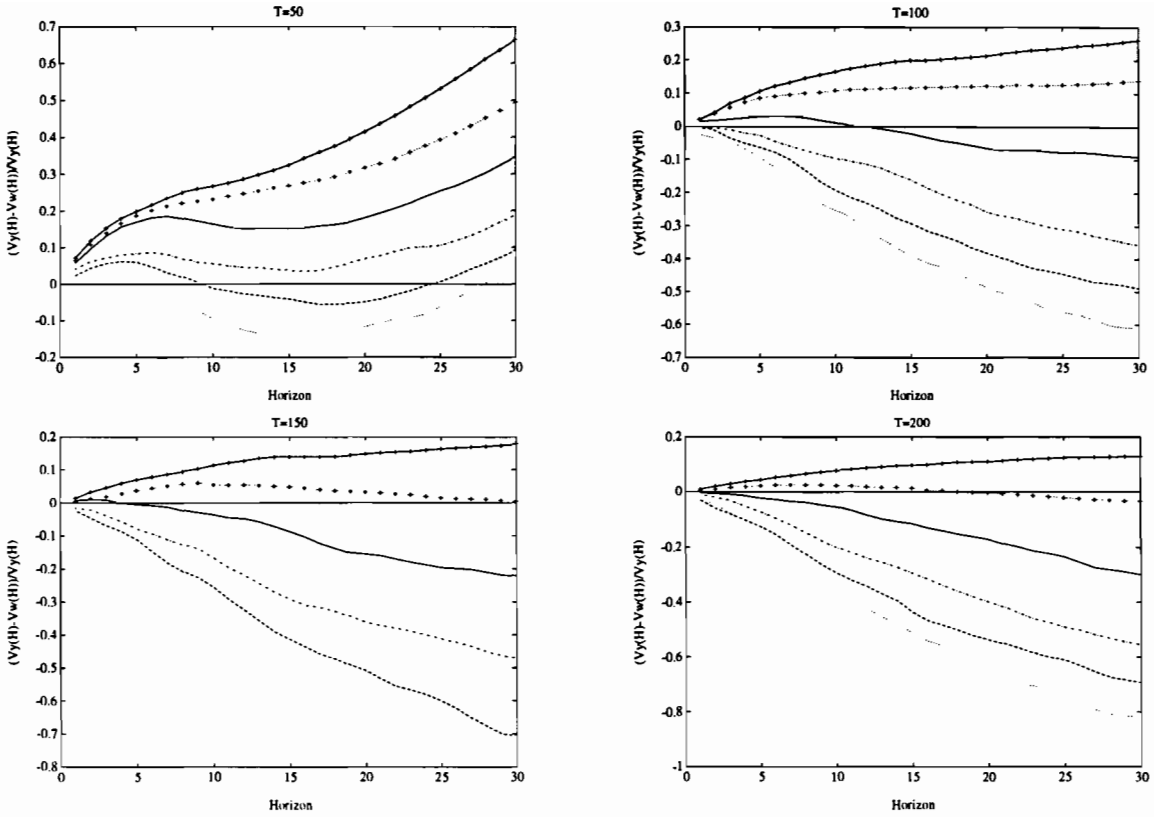

Figure 2: $(V_y(H) - V_w(H))/V_y(H)$ of model M1 for horizon $H = 1, .., 30$ and sample size T. The values of ρ are (from down to top): 0.90, 0.92, 0.94, 0.96, 0.98 , 0.99.

advantages of parsimony in order to obtain accurate forecasts. As described in the preceding subsection, both table (6.1) and expression (6.15) show that, as the horizon increases, it is neccesary to be closer to nonstationarity to hold (6.14) although this variation is small.

7 A simulation study

In this section we present a simulation exercise to illustrate the preceding results and validate them in finite samples. Three different AR(2) were analyzed,

$$\text{M1: } (1 - 0.5B)(1 - \rho B)y_t = 10 + a_t$$

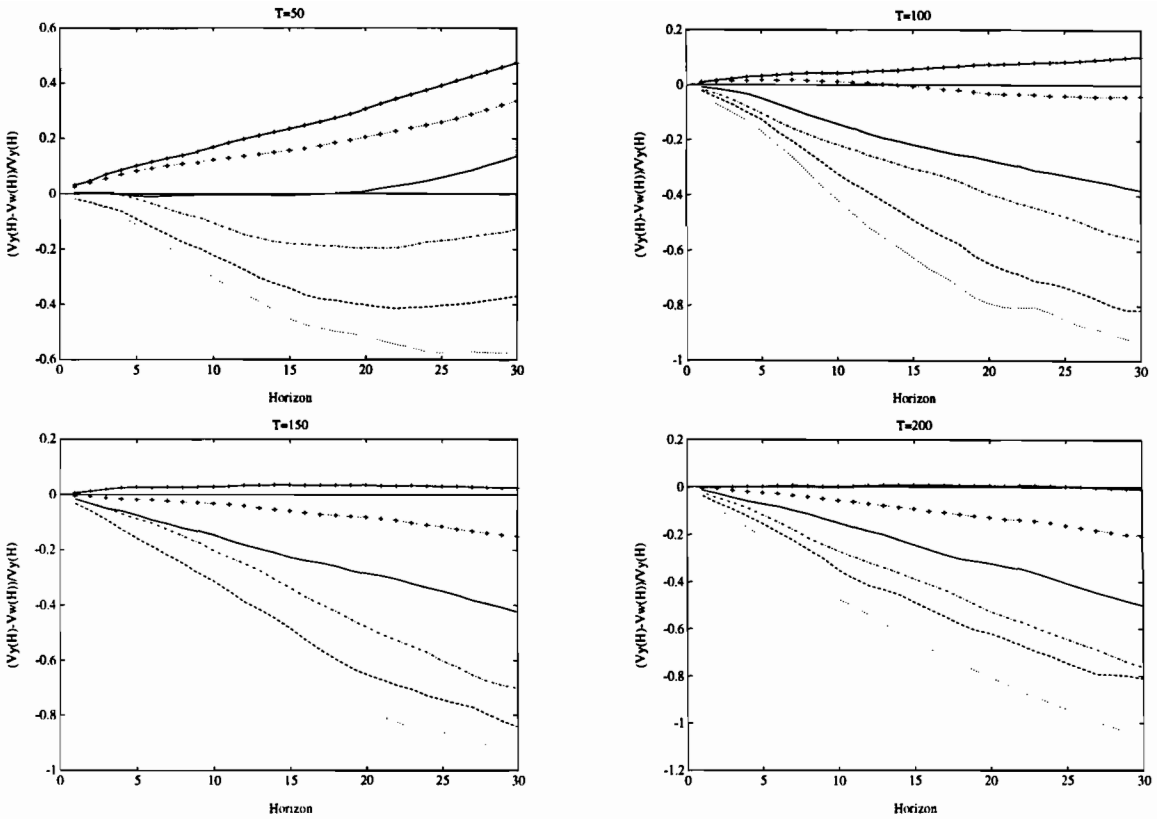

Figure 3: $(V_y(H) - V_w(H))/V_y(H)$ of model M2 for horizon $H = 1, .., 30$ and sample size T. The values of ρ are (from down to top): 0.90, 0.92, 0.94, 0.96, 0.98 , 0.99.

$$\text{M2: } (1 - 0.5B)(1 - \rho B)y_t = a_t$$

$$\text{M3: } (1 + 0.8B)(1 - \rho B)y_t = 10 + a_t.$$

with $\rho = 0.9,\ 0.92,\ 0.94,\ 0.96,\ 0.98,\ 0.99$, and four samples sizes $T = 50,\ 100,\ 150,\ 200$. Real series usually have non zero mean and models M1 and M3 can illustrate the consequences of overdifferencing in such series. Nevertheless when the decision is on taking the second difference we should expect a situation with zero mean as in model M2.

In each replication we generate a random sample of the process of size $500 + T + 30$ with noise $a_t \sim N(0, 1)$. The first 500 observations were dropped to avoid the effect of initial values and the last 30 were reserved to evaluate the prediction error. By averaging the

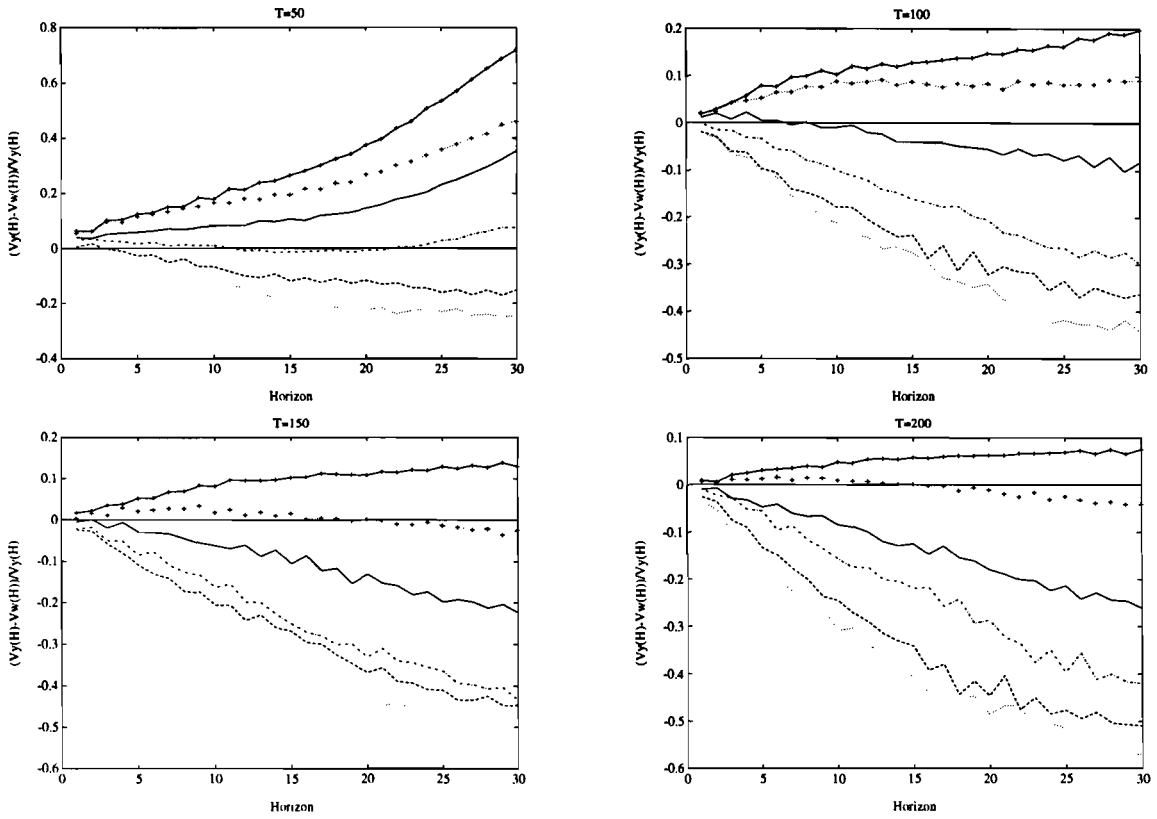

Figure 4: $(V_y(H) - V_w(H))/V_y(H)$ of model M3 for horizon $H = 1, .., 30$ and sample size T. The values of ρ are (from down to top): 0.90, 0.92, 0.94, 0.96, 0.98 , 0.99.

predicting square errors of 5000 replications we obtain $V_y(H)$ and $V_w(H)$ as the sampling estimation of the PMSE of forecasting y_{T+H} using the forecasts generated by the correct AR(2) model or the overdifferenced ARIMA(1,1,0) model respectively. Figures 2 to 4 shows the ratio $(V_y(H) - V_w(H))/V_y(H)$ for M1 to M3 and differents values of T and ρ. Such a ratio represents the expected gain (or loss if negative) of overdifferencing at each horizon.

These figures reveals that, as expected from the theoretical results, there are situations where overdifferencing overpass the true model. The expected gain increases with the size of ρ and decreases with T. Also, congruently with equation (3.4), the gain is larger in the model with positive second root (M1) than in the model with negative root (M3). The gain substantially decreases if $\alpha = 0$ (M2).

The main feature of these figures is the divergence of the curves as the horizon increases. At $H = 1$ the difference of the alternating modellizations is very small, even negligible. However at long term the gain or loss is rather important in most of the cases. A second feature is the asymetry of the consequences of overdifferencing. In large sample sizes the loss of forecast efficiency can be severe if we overdifferentiate and ρ is not large enough, whereas the gain, in the case of largest root very close to unity, is modest; conversely, in the case of small sample sizes, the gain of overdifferencing is substancial whereas the loss if ρ is not large enough can be modest. This effect was already proven in theorems 7 and 8. If ρ is close enough to one ($\beta > 1$) the PMSE of the overdifferenced model is lower than in the true model by a factor of magnitude $O(T^{-1})$ and the gain will be also $O(T^{-1})$, whereas if $\beta < 1$ the overdifferenced predictor increases its PMSE by a factor bigger than $O(T^{-1})$ and therefore the loss of forecast efficiency is larger than $O(T^{-1})$.

It can also be observed that there are cases where the sign of the ratio changes. For instance, in Figure 2, if $T = 150$ and $\rho = 0.96$ there is a small positive expected gain if overdifferencing but a expected loss of 20% in 24 steps ahead forecasts. This effect was already detected in section 6 where, for longer horizons, larger values of $\phi(h)$ were needed in order to expect a positive gain in overdifferencing. This can be seen in Figures 2 to 4 as an added convexity to the smooth growing shape of the positive gain that could lead to reach the negative zone in some cases. With $T = 50$ the sign can even change twice, but this second change happens at very high H compared with the sample size and also a extremely high variability is expected.

Since results depend mainly on the size of the roots rather than in its number it is reasonable to foresee similar conclusions in larger autorregresions than used in these simulations.

8 Summary and concluding remarks

We have proved in this article that, when the largest root of an AR($p + 1$) process is $\rho = exp(-c/T^{\beta})$, $c > 0$, $\beta > 1$, the estimation of a misspecified overdifferenced ARIMA($p, 1, 0$) is root-T consistent and its PMSE is lower than in the correct predictor due to its parsimony. This superiority of the overdifferenced predictor is small in the short term but increases, in general, in the long term. However, when such a largest root is

not as close to unity, the misspecified predictor can still have similar one-step ahead forecasting performance than the correct model, but long term forecasting are seriuosly afected.

Our results also diminish the importance of the low power of unit roots tests in the region of near nonstationarity. Although in this case the test could lead to fit a misspecified ARIMA$(p, 1, 0)$, this model will provide better forecast performance than the correct model.

As seen in the previous section, there are situations where the optimal model to predict one period ahead is not necesarily the optimal at longer horizons. In some other situations the optimality for short term forecasting is hardly determined because the competing predictors are very similar, whereas at longer horizons the optimal predictor is clear. This different behaviour of the relative forecasting performance at the short and long term reveals that diagnosis based on one step ahead forecasting performance could be a poor, even misleading, tool to choose the optimal predictor in the boundary of nonstationarity. Therefore, if a predictor is gone to be used to forecast H steps ahead, it is advisable to analyze the properties of the candidate predictors in such a horizon. The idea of using different models for multi-period forecasting independently of which is the correct one has largely been used in statistical modelling (Cox 1961, Gersh & Kitagawa 1983, Findley 1984, Weiss 1991, Tsay 1993, Tiao & Tsay 1994, Tiao & Xu 1993 among others). These authors justify this practice under the assumption that the choosen models are wrong and therefore *the choice of the best approximate model would depend on the particular purpose of the model* (Gersh & Kitagawa 1983, p. 262). In our case this recomendation still holds even though the correct model were known but not its parameters. The sampling variability due to parameter estimation makes that the estimated correct model can be surpassed in some limiting situations, as near nonstationarity. Better predictions, specially at long term, can be obtained by using a more parsimonious model than the correct one.

Acknowledgements

Both authors acknowledge support for this research provided by DGICYT (Spain) under grant PB93-0232.

References

Berk, K. N. (1974). "Consistent Autoregresive Spectral Estimates," *The Annals of Statistics,* 2, 489–502.

Bhansali, R.J. (1978). "Linear Prediction by Autoregressive Model Fitting in the Time Domain," *The Annals of Statistics,* 6, 224–231.

Bhansali, R. J. (1981). "Effects of Not Knowing the Order of an Autoregressive Process on the Mean Squared Error of Prediction– I," *Journal of the American Statistical Association,* 76, 588–597.

Box, G. E. P. and Tiao, G. C. (1977). "A Canonical Analysis of Multiple Time Series," *Biometrika,* 6, 355–365.

Cox, D. R. (1961). "Prediction by Exponentially Weighted Moving Averages and Related Methods," *Journal of the Royal Statistical Society, Serie B,* 23, 414–422.

Davies, N. and Newbold, P. (1980). "Forecasting with Misspecified Models," *Applied Statistics,* 29, 87–92.

Dickey, D. A., and Fuller, W. A. (1979). "Distribution of the Estimators for Autoregresive Time Series With a Unit Root," *Journal of the American Statistical Association,* 74, 427–431.

Fuller, W. A. (1976), *Introduction to Statistical Time Series.* New York: John Wiley.

Findley, D. F. (1984). "On Some Ambiguities Associated With the Fitting of ARMA Models to Time Series," *Journal of Time Series Analysis,* 5, 213–225.

Gersh, W. and Kitagawa, G. (1983). "The Prediction of Time Series With Trends and Seasonalities," *Journal of Business & Economic Statistics,* 1, 253–263.

Harvey, A. C. (1981), "Finite Sample Prediction and Overdifferencing," *Journal of Time Series Analysis,* 2, 221– 232.

Hosoya, Y. and Taniguchi, M. (1982). "A Central Limit Theorem for Stationary Processes and the Parameter Estimation of Linear Processes," *The Annals of Statistics,* 10, 132–153.

Kendall, M. G. (1954). "Note on Bias in the Estimation of Autocorrelation," *Biometrika,* 41, 403–404.

Kunitomo, N. and Yamamoto, T. (1985). "Properties of Predictors in Misspecified Autoregressive Time Series Models," *Journal of the American Statistical Association,* 80, 941–950.

Ledolter, J. and Abraham, B. (1981). "Parsimony and Its Importance in Time Series Forecasting," *Technometrics,* 23, 411–414.

Lütkepohl, H. (1991). *Introduction to Multiple Time Series Analysis.* Berlin: Springer-Verlag.

Marriot, F. H. C. and Pope, J. A. (1954). "Bias in Estimation of Autocorrelations," *Biometrika,* 41, 390–402.

Phillips, P. C. B. (1987). "Towards a Unified Asymptotic Theory for Autoregression," *Biometrika,* 74, 535–547.

Plosser, C. I. and Schwert, G. W. (1977), "Estimation of a Non-invertible Moving Average Process: The Case of Overdifferencing," *Journal of Econometrics,* 6, 199–224.

Plosser, C. I. and Schwert, G. W. (1978), "Money, Income and Sunspots: Measuring Economic Relationships and the Effects of Differencing," *Journal of Monetary Economics,* 4, 637–660.

Saikkonen, P. and Luukkonen, R. (1993). "Testing for a Moving Average Unit Root in Autoregressive Integrated Moving Average Models," *Journal of the American Statistical Association,* 88, 596–601

Tanaka, K. (1984). "An Asymptotic Expansion Associated with the Maximum Likelihood Estimators in ARMA Models," *Journal of the Royal Statistical Society, Serie B,* 46, 58–67.

Tanaka, K. (1990). "Testing for a Moving Average Unit Root," *Econometric Theory,* 6, 433–444.

Tanaka, K. and Maekawa, K. (1984). "The Sampling Distributions of the Predictor for an Autoregression Model under Misspecifications," *Journal of Econometrics,* 25, 327–351.

Tiao G. C. and Tsay, R. S. (1994). "Some Advances in Nonlinear and Adaptive Modelling in Time Series Analysis," *Journal of Forecasting,* 13, 109–131

Tiao, G. C. and Xu, D. (1993). "Robustness of Maximum Likelihood Estimates for Multi-step Predictions: The Exponential Smoothing Case," *Biometrika,* 80, 623–641.

Tjøsteim, D. and Paulsen, J. (1983). "Bias of Some Commonly-Used Time Series Estimates," *Biometrika,* 70,, 389–399.

Tsay, R. S. (1993). "Adaptive Forecasting, A Comment on 'Calculating Interval Forecasts' by Chatfield, C.," *Journal of Business & Economic Statistics,* 11, 140–142.

Tsay, R. S. (1993), "Testing for Noninvertible Models With Applications," *Journal of Business & Economic Statistics,* 11, 225–233.

Shaman, P. and Stine, R. (1988). "The Bias of Autoregressive Coefficient Estimators," *Journal of the American Statistical Association,* 83, 842–848.

Shenton, L. R. and Johnson, W. L. (1965). "Moments of a Serial Correlation Coefficient," *Journal of the Royal Statistical Society, Serie B,* 27, 308–320.

Weiss, A. A. (1991). "Multi-step Estimation and Forecasting in Dynamic Models," *Journal of Econometrics,* 48, 135–149.

White, J. S. (1961). "Asymptotic Expansions for the Mean and Variance of Serial Correlation Coefficient," *Biometrika,* 48, 85–94.

APPENDIX

A Some previous lemmas

We present some lemmas for the proof of theorems in subsequent sections. For an arbitrary $r \times 1$ vector x and a $r \times r$ matrix M, let $\|x\| = (x'x)^{1/2}$ be the Euclidean norm of x and $\|M\| = sup_{\|x\| \leq 1}(x'M'Mx)^{1/2}$ be the matrix norm of M.

Lemma A.1 *Assume A1 and A2, with $s_o = 2k$ and $k \geq 1$. Then as $T \to \infty$*

$$E(\|\hat{\Gamma}_w - \hat{\Gamma}_{w|p}\|^k) = O\left(\left[\frac{1-\rho}{1+\rho}\right]^{\frac{k}{2}}\right), \tag{A.1}$$

and

$$E(\|\hat{\gamma}_w - \hat{\gamma}_{w|p}\|^k) = O\left(\left[\frac{1-\rho}{1+\rho}\right]^{\frac{k}{2}}\right). \tag{A.2}$$

Proof: The matrix norm is of the same magnitude order than the norm of its largest element. Then $E(\|\hat{\Gamma}_w - \hat{\Gamma}_{w|p}\|^k) = O\left(E\{[(w_t w_{t-s} - w_{t|p}w_{t-s|p})^2]^{\frac{k}{2}}\}\right)$, and similar result applies to (A.2). Using the decomposition (3.6) we have

$$\begin{aligned} w_t w_{t-s} - w_{t|p}w_{t-s|p} &= -\sum_{i=0}^{\infty}\psi_i(1-\rho)z_{t-s-1-i}w_{t|p} - \sum_{j=0}^{\infty}\psi_j(1-\rho)z_{t-1-j}w_{t-s|p} \\ &\quad + \sum_{j=0}^{\infty}\sum_{i=0}^{\infty}\psi_j\psi_i(1-\rho)^2 z_{t-1-j}z_{t-s-1-i} \\ &= -a - b + c, \end{aligned}$$

where $a = \sum_{i=0}^{\infty}\psi_i(1-\rho)z_{t-s-1-i}w_{t|p}$, $b = \sum_{j=0}^{\infty}\psi_j(1-\rho)z_{t-1-j}w_{t-s|p}$ and $c = \sum_{j=0}^{\infty}\sum_{i=0}^{\infty}\psi_j\psi_i(1-\rho)^2 z_{t-1-j}z_{t-s-1-i}$. By Minkowski's inequality

$$E|-a-b+c|^k \leq \left\{\left(E|a|^k\right)^{\frac{1}{k}} + \left(E|b|^k\right)^{\frac{1}{k}} + \left(E|c|^k\right)^{\frac{1}{k}}\right\}^k.$$

Let us solve the term $E|a|^k$:

$$\begin{aligned} E|a|^k &= E\left[\left|\sum_{i=0}^{\infty}\psi_i(1-\rho)z_{t-s-1-i}w_{t|p}\right|^k\right] \\ \sum_{i_1=0}^{\infty}\cdots\sum_{i_k=0}^{\infty}&\left|\psi_{i_1}\cdots\psi_{i_k}(1-\rho)^k\right| \times E[|z_{t-s-1-i_1}\cdots z_{t-s-1-i_k}w_{t|p}^k|], \end{aligned} \tag{A.3}$$

where

$$E[|z_{t-s-1-i_1}\cdots z_{t-s-1-i_k}w_{t|p}^k|] \leq \sum_{r_1=0}^{\infty}\cdots\sum_{r_k=0}^{\infty}\sum_{m_1=0}^{\infty}\cdots\sum_{m_k=0}^{\infty}\rho^{r_1}\cdots\rho^{r_k}|\psi_{m_1}\cdots\psi_{m_k}|$$
$$\times E(|a_{t-(r_1+i_1+s+1)}\cdots a_{t-(r_k+i_k+s+1)}a_{t-m_1}\cdots a_{t-m_k}|). \quad (A.4)$$

By assumption A1 we have $E(|a_t|^{s_0}) < \infty$. Therefore the term in the last line of (A.4) is $O(1)$ if the a_t's match pairwise and is null otherwise. To simplify the proof let us consider the case $k = 2$, then

$$E[|z_{t-s-1-i_1}z_{t-s-1-i_2}w_{t|p}^2|] \leq \sum_{r_1=0}^{\infty}\sum_{r_2=0}^{\infty}\sum_{m_1=0}^{\infty}\sum_{m_2=0}^{\infty}\rho^{r_1}\rho^{r_2}|\psi_{m_1}\psi_{m_2}|$$
$$\times E(|a_{t-(r_1+i_1+s+1)}a_{t-(r_2+i_2+s+1)}a_{t-m_1}a_{t-m_2}|). \quad (A.5)$$

Let us denote as $\mathcal{S}_1$ the situations where the four a_t's in (A.5) match. Then $\{(r_1+i_1+s+1) = (r_2+i_2+s+2) = m_1 = m_2\}$; and hence, assuming without lost of generality the case $i_1 > i_2$

$$E[|z_{t-s-1-i_1}z_{t-s-1-i_2}w_{t|p}^2| \mid \mathcal{S}_1] = \sum_{r_1=0}^{\infty}\rho^{r_1}\rho^{r_1+i_1+2s+2-i_2}\psi_{r_1+i_1+s+1}^2 O(1)$$
$$\leq \sum_{r_1=0}^{\infty}\psi_{r_1}^2 O(1) = O(1).$$

A second possibility is that the a_t's in (A.5) match in pairs but each pair contains the cross product of z's and w's simultaneously as in the case $\{(r_1 + i_1 + s + 1 = m_1) \neq (r_2 + i_2 + s + 2 = m_2)\}$. Let us denote this situation as $\mathcal{S}_2$, then

$$E[|z_{t-s-1-i_1}z_{t-s-1-i_2}w_{t|p}^2| \mid \mathcal{S}_2] = \left(\sum_{r_1=0}^{\infty}\rho^{r_1}|\psi_{r_1+i_1+s+1}|O(1)\right)\left(\sum_{r_2=0}^{\infty}\rho^{r_2}|\psi_{r_2+i_2+s+1}|O(1)\right)$$
$$\leq \left(\sum_{r_1=0}^{\infty}|\psi_{r_1}|\right)^2 = O(1).$$

The third possibility is that the a_t's in (A.5) match in pairs but one pair contains the product of z's and the other of w's, then $\{(r_1 + i_1 + s + 1 = r_2 + i_2 + s + 2) \neq (m_1 = m_2)\}$ then, assuming without lost of generality the case $i_1 > i_2$ and denoting as $\mathcal{S}_3$ to this situation

$$E[z_{t-s-1-i_1}z_{t-s-1-i_2}w_{t|p}^2 \mid \mathcal{S}_3] = \left(\sum_{r_1=0}^{\infty}\rho^{r_1}\rho^{r_1+i_1+2s+2-i_2}O(1)\right)\left(\sum_{m_1=0}^{\infty}\psi_{m_1}^2 O(1)\right)$$

$$\leq \left(\sum_{r_1=0}^{\infty} \rho^{2r_1}\right)\left(\sum_{m_1=0}^{\infty} \psi_{m_1}^2\right) O(1) = O\left(\frac{1}{1-\rho^2}\right).$$

Therefore this third case supplies the terms of largest magnitude. Hence, in the general case, we only will be interested in the situations where the a_t's from z's match pairwise but do not match with the a_t's of w's. If k is even and using that

$$\frac{1}{1-\rho^k} < \left(\frac{1}{1-\rho^{k-2}}\right)\left(\frac{1}{1-\rho^2}\right) < \cdots < \frac{1}{(1-\rho^2)^{\frac{k}{2}}},$$

it can be verified that the cases of larger magnitude are those where the a_t's from z's match pairwise but with no matching between pairs. Then

$$E[|z_{t-s-1-i_1} \cdots z_{t-s-1-i_k} w_{t|p}^k|] = O\left(\left[\frac{1}{1-\rho^2}\right]^{\frac{k}{2}}\right).$$

Similarly, if k is odd, the largest terms will be produced by matching pairwise the a_t's from z's, but in this case we only have $k-1$ pairs. Then

$$E[|z_{t-s-1-i_1} \cdots z_{t-s-1-i_k} w_{t|p}^k|] = O\left(\left[\frac{1}{1-\rho^2}\right]^{\frac{k-1}{2}}\right).$$

Applying this results to (A.3) we obtain

$$E|a|^k \leq \sum_{i_1=0}^{\infty} \cdots \sum_{i_k=0}^{\infty} \left|\psi_{i_1} \cdots \psi_{i_k} (1-\rho)^k\right| O\left(\left[\frac{1}{1-\rho^2}\right]^{\frac{k}{2}}\right) = O\left(\left[\frac{1-\rho}{1+\rho}\right]^{\frac{k}{2}}\right).$$

It is easily verified, following the previous arguments, that

$$E|b|^k = O(E|a|^k) = O\left(\left[\frac{1-\rho}{1+\rho}\right]^{\frac{k}{2}}\right).$$

In the sane way we can solve the term $E|c|^k$.

$$\begin{aligned} E|c|^k &= E\left[\left|\sum_{j=0}^{\infty}\sum_{i=0}^{\infty} \psi_j \psi_i (1-\rho)^2 z_{t-1-j} z_{t-s-1-i}\right|^k\right] \\ &\leq \sum_{j_1=0}^{\infty} \cdots \sum_{j_k=0}^{\infty}\sum_{i_1=0}^{\infty} \cdots \sum_{i_k=0}^{\infty} |\psi_{j_1} \cdots \psi_{j_k} \psi_{i_1} \cdots \psi_{i_k} (1-\rho)^{2k}| \\ &\quad \times E(|z_{t-1-j_1} \cdots z_{t-1-j_k} z_{t-s-1-i_1} \cdots z_{t-s-1-i_k}|), \end{aligned} \tag{A.6}$$

where

$$E(|z_{t-1-j_1}\cdots z_{t-1-j_k}z_{t-s-1-i_1}\cdots z_{t-s-1-i_k}|) \leq \sum_{r_1=0}^{\infty}\cdots\sum_{r_k=0}^{\infty}\sum_{m_1=0}^{\infty}\cdots\sum_{m_k=0}^{\infty}\rho^{r_1}\cdots\rho^{r_k}\rho^{m_1}\cdots\rho^{m_k}$$
$$\times E(|a_{t-(r_1+j_1+1)}\cdots a_{t-(r_k+j_1+1)}a_{t-(m_1+i_1+s+1)}\cdots a_{t-(m_k+i_k+s+1)}|). \quad (A.7)$$

Similarly to the arguments exposed in the analysis of $E(|a|^k)$, the magnitude of (A.7) is determined by the terms with a_t's matching pairwise but with no matching between pairs. If k is even the magnitude of these terms is $O((1-\rho^2)^{-k})$, whereas if k is odd the magnitude is $O((1-\rho^2)^{-(k-1)})$. Since there is a finite number of such terms we obtain from (A.6)

$$E(|c|^k) = O\left(\frac{(1-\rho)^{2k}}{(1-\rho^2)^k}\right) = O\left(\left[\frac{1-\rho}{1+\rho}\right]^k\right).$$

Since the magnitude order when k is odd is smaller than in the case of k even, we express all the situations as if it were even, to avoid further complexity in notation. Therefore we conclude that

$$E|-a-b+c|^k = O\left(\left[\frac{1-\rho}{1+\rho}\right]^{-\frac{k}{2}}\right),$$

and the lemma holds. □

Lemma A.2 *Assume A1, A2, A3 and A3', with $s_0 = 2k$. Then as $T \to \infty$*

$$E(\|\hat{\Gamma}_w^{-1} - \hat{\Gamma}_{w|p}^{-1}\|^k) = O\left(\left[\frac{1-\rho}{1+\rho}\right]^{\frac{k}{2}}\right).$$

Proof: It can be verified that

$$\|\hat{\Gamma}_w^{-1} - \hat{\Gamma}_{w|p}^{-1}\|^k = \|\hat{\Gamma}_w^{-1}(\hat{\Gamma}_w - \hat{\Gamma}_{w|p})\hat{\Gamma}_{w|p}^{-1}\|^k \leq \|\hat{\Gamma}_w^{-1}\|^k\|\hat{\Gamma}_w - \hat{\Gamma}_{w|p}\|^k\|\hat{\Gamma}_{w|p}^{-1}\|^k.$$

By Hölders' inequality and lemma A.1,

$$\begin{aligned} E(\|\hat{\Gamma}_w^{-1} - \hat{\Gamma}_{w|p}^{-1}\|^k) &\leq \left(E(\|\hat{\Gamma}_w^{-1}\|^{4k})E(\|\hat{\Gamma}_{w|p}^{-1}\|^{4k})\right)^{\frac{1}{4}}\left(E(\|\hat{\Gamma}_w - \hat{\Gamma}_{w|p}\|^{2k})\right)^{\frac{1}{2}} \\ &= O\left(\left[\frac{1-\rho}{1+\rho}\right]^{\frac{k}{2}}\right), \end{aligned}$$

and the lemma hold. □

Lemma A.3 *Assume A1, A2, A3 and A3', with* $s_0 = 2k$. *Then as* $T \to \infty$

$$E(\|\hat{\phi} - \hat{\phi}_{|p}\|^k) = O\left(\left[\frac{1-\rho}{1+\rho}\right]^{\frac{k}{2}}\right), \tag{A.8}$$

and if $s_0 = 4k$

$$E(\|\hat{\phi} - \phi\|^k) = O\left(max\left\{\left[\frac{1-\rho}{1+\rho}\right]^{\frac{k}{2}}, T^{-\frac{k}{2}}\right\}\right). \tag{A.9}$$

Proof: The estimator $\hat{\phi}$ can be expressed as

$$\begin{aligned}\hat{\phi} &= \hat{\Gamma}_w^{-1}\hat{\gamma}_w = (\hat{\Gamma}_w^{-1} - \hat{\Gamma}_p^{-1} + \hat{\Gamma}_p^{-1})(\hat{\gamma}_w - \hat{\gamma}_p + \hat{\gamma}_p) \\ &= (\hat{\Gamma}_w^{-1} - \hat{\Gamma}_p^{-1})(\hat{\gamma}_w - \hat{\gamma}_{w|p}) + (\hat{\Gamma}_w^{-1} - \hat{\Gamma}_{w|p}^{-1})\hat{\gamma}_{w|p} + \hat{\Gamma}_{w|p}^{-1}(\hat{\gamma}_w - \hat{\gamma}_{w|p}) + \hat{\phi}_{|p},\end{aligned}$$

where $\hat{\phi}_{|p} = \hat{\Gamma}_{w|p}^{-1}\hat{\gamma}_{w|p}$. By the c_r inequality we obtain

$$\begin{aligned}E(\|\hat{\phi} - \hat{\phi}_{w|p}\|^k) \;\leq\; & 2^{k-1}E[(\|\hat{\Gamma}_w^{-1} - \hat{\Gamma}_{w|p}^{-1})(\hat{\gamma}_w - \hat{\gamma}_{w|p})\|] \\ & +2^{k-1}\left\{2^{k-1}E[\|(\hat{\Gamma}_w^{-1} - \hat{\Gamma}_{w|p}^{-1})\hat{\gamma}_{w|p}\|] + 2^{k-1}E[\|\hat{\Gamma}_{w|p}^{-1}(\hat{\gamma}_w - \hat{\gamma}_{w|p})\|]\right\}.\end{aligned}$$

By Hölder inequality and applying lemma A.2 we have

$$E\left[\|(\hat{\Gamma}_w^{-1} - \hat{\Gamma}_{w|p}^{-1})(\hat{\gamma}_w - \hat{\gamma}_{w|p})\|^k\right] \leq \left\{E(\|\hat{\Gamma}_w^{-1} - \hat{\Gamma}_{w|p}^{-1}\|^{2k})E(\|\hat{\gamma}_w - \hat{\gamma}_{w|p}\|^{2k})\right\}^{\frac{1}{2}} = O\left(\left[\frac{1-\rho}{1+\rho}\right]^k\right),$$

and also

$$E[\|(\hat{\Gamma}_w^{-1} - \hat{\Gamma}_{w|p}^{-1})\hat{\gamma}_{w|p}\|^k] \leq \left\{E(\|\hat{\Gamma}_w^{-1} - \hat{\Gamma}_{w|p}^{-1}\|^{2k})E(\|\hat{\gamma}_{w|p}\|^{2k})\right\}^{\frac{1}{2}} = O\left(\left[\frac{1-\rho}{1+\rho}\right]^{\frac{k}{2}}\right).$$

Applying assumption A3 we have

$$E[\|\hat{\Gamma}_{w|p}^{-1}(\hat{\gamma}_w - \hat{\gamma}_{w|p})\|^k] \leq \left\{E(\|\hat{\Gamma}_{w|p}^{-1}\|^{2k})E(\|\hat{\gamma}_w - \hat{\gamma}_p\|^{2k})\right\}^{\frac{1}{2}} = O\left(\left[\frac{1-\rho}{1+\rho}\right]^{\frac{k}{2}}\right),$$

and then (A.8) holds. To proof (A.9) we will use the decomposition

$$\hat{\phi} - \phi = \hat{\Gamma}_w^{-1}\hat{\gamma}_w - \Gamma_{w|p}^{-1}\gamma_{w|p} = \Gamma_{w|p}^{-1}(\hat{\gamma}_w - \gamma_{w|p}) + \hat{\gamma}_w(\hat{\Gamma}_w^{-1} - \Gamma_{w|p}^{-1}).$$

Then, by the c_r and Hölder inequalities

$$E[\|\hat{\phi} - \phi\|^k] \leq 2^{k-1}\left\{E[\|\Gamma_{w|p}^{-1}\|^{2k}E[\|\hat{\gamma}_w - \gamma_{w|p}\|^{2k}]\right\}^{\frac{1}{2}} + 2^{k-1}\left\{E[\|\hat{\gamma}_w\|^{2k}]E[\|\hat{\Gamma}_w^{-1} - \Gamma_{w|p}\|^{2k}]\right\}^{\frac{1}{2}},$$

where, by lemma A.1 and using lemma 3.3 of Bhansali (1981) (with $s_0 = 4k$),

$$\begin{aligned} E[\|\hat{\gamma}_w - \gamma_{w|p}\|^k] &\leq 2^{k-1}E[\|\hat{\gamma}_w - \hat{\gamma}_{w|p}\|^k] + 2^{k-1}E[\|\hat{\gamma}_{w|p} - \gamma_{w|p}\|^k] \\ &= O\left(max\left\{\left[\frac{1-\rho}{1+\rho}\right]^{\frac{k}{2}}, T^{-\frac{k}{2}}\right\}\right), \end{aligned}$$

and, by lemma A.2,

$$\begin{aligned} E[\|\hat{\Gamma}_w^{-1} - \Gamma_{w|p}^{-1}\|^k] &\leq 2^{k-1}E[\|\hat{\Gamma}_w^{-1} - \hat{\gamma}_{w|p}^{-1}\|^k] + 2^{k-1}E[\|\hat{\Gamma}_{w|p}^{-1} - \Gamma_{w|p}^{-1}\|^k] \\ &= O\left(max\left\{\left[\frac{1-\rho}{1+\rho}\right]^{\frac{k}{2}}, T^{-\frac{k}{2}}\right\}\right). \end{aligned}$$

□

B Proofs of section 3

Proof of theorem 2: By (2.7)

$$\hat{\phi} - \phi = \hat{\Gamma}_w^{-1}\hat{\gamma}_w - \Gamma_{w|p}^{-1}\gamma_{w|p} = \Gamma_{w|p}^{-1}(\hat{\gamma}_w - \gamma_{w|p}) + \hat{\gamma}_w(\hat{\Gamma}_w^{-1} - \Gamma_{w|p}^{-1}).$$

By stationarity of $\{w_{t|p}\}$ we have $\Gamma_{w|p}^{-1} = O(1)$. From corollary 1 $(\hat{\gamma}_w - \gamma_{w|p}) = O_p(T^{-1/2})$. Also, if $\hat{\Gamma}_w^{-1}$ exist, we have $(\hat{\Gamma}_w^{-1} - \Gamma_{w|p}^{-1}) = \hat{\Gamma}_w^{-1}(\Gamma_{w|p} - \hat{\Gamma}_{w|p}^{-1})\hat{\Gamma}_{w|p}^{-1} = O_p(T^{-1/2})$, where corollary 1 has been applied. Therefore $\hat{\phi} - \phi = O_p(T^{-1/2})$. □

Proof of theorem 3: The estimator $\hat{\phi}$ can be expressed as

$$\begin{aligned} \hat{\phi} &= \hat{\Gamma}_w^{-1}\hat{\gamma}_w = (\hat{\Gamma}_w^{-1} - \hat{\Gamma}_{w|p}^{-1} + \hat{\Gamma}_{w|p}^{-1})(\hat{\gamma}_w - \hat{\gamma}_{w|p} + \hat{\gamma}_{w|p}) \\ &= (\hat{\Gamma}_w^{-1} - \hat{\Gamma}_{w|p}^{-1})(\hat{\gamma}_w - \hat{\gamma}_{w|p}) + (\hat{\Gamma}_w^{-1} - \hat{\Gamma}_{w|p}^{-1})\hat{\gamma}_{w|p} + \hat{\Gamma}_{w|p}^{-1}(\hat{\gamma}_w - \hat{\gamma}_{w|p}) + \hat{\phi}_{|p}, \end{aligned}$$

where $\hat{\phi}_{|p} = \hat{\Gamma}_{w|p}^{-1}\hat{\gamma}_{w|p}$. Hence $E(\hat{\phi}) = E(\hat{\phi}_{|p}) + O(R)$. By Hölders' inequality and applying lemma A.2 and A.3

$$O(\|R\|) = O\left(\left[\frac{1-\rho}{1+\rho}\right]^{\frac{1}{2}}\right),$$

and

$$E(\hat{\phi} - \phi) = E(\hat{\phi}_{|p} - \phi) + O\left(\left[\frac{1-\rho}{1+\rho}\right]^{\frac{1}{2}}\right). \tag{B.1}$$

By Bhansali (1981), $E(\hat{\phi}_{|p} - \phi) = O(T^{-1})$ (with $s_0 = 16$). □

Proof of theorem 4: We can decompose

$$\begin{aligned}(\hat{\phi} - \phi)(\hat{\phi} - \phi)' &= (\hat{\phi} - \hat{\phi}_{|p})(\hat{\phi} - \hat{\phi}_{|p})' + 2(\hat{\phi} - \phi)(\hat{\phi}_{|p} - \phi)'] \\ &\quad +(\hat{\phi}_{|p} - \phi)(\hat{\phi}_{|p} - \phi)'.\end{aligned}$$

Therefore $E[(\hat{\phi} - \phi)(\hat{\phi} - \phi)'] = E[(\hat{\phi}_{|p} - \phi)(\hat{\phi}_{|p} - \phi)'] + O(R)$, where

$$R = E[(\hat{\phi} - \hat{\phi}_{|p})(\hat{\phi} - \hat{\phi}_{|p})'] + 2E[(\hat{\phi} - \phi)(\hat{\phi}_{|p} - \phi)'].$$

Using that the 2-norm of a matrix M is the largest singular value of $M'M$ then $\|M\| \leq \sqrt{\text{trace}(M'M)}$. Hence

$$\begin{aligned}E[\|(\hat{\phi} - \hat{\phi}_{|p})(\hat{\phi} - \hat{\phi}_{|p})'\|] &\leq E\left[\text{trace}\left((\hat{\phi} - \hat{\phi}_{|p})(\hat{\phi} - \hat{\phi}_{|p})'(\hat{\phi} - \hat{\phi}_{|p})(\hat{\phi} - \hat{\phi}_{|p})'\right)^{\frac{1}{2}}\right] \\ &= E[\|\hat{\phi} - \hat{\phi}_{|p}\|^2].\end{aligned}$$

Similarly, and using Hölders' inequality

$$E[\|(\hat{\phi} - \phi)(\hat{\phi}_{|p} - \phi)\|] \leq \left\{E[\|(\hat{\phi} - \phi)\|^2]E[\|(\hat{\phi}_{|p} - \phi)\|^2]\right\}^{\frac{1}{2}}.$$

Since $E(\|\hat{\phi}_{|p} - \phi\|^2) = O(T^{-1})$ (see, for instance, Bhansali 1981) and applying lemma A.3 we obtain

$$O(\|R\|) = O\left(\max\left\{\left[\frac{1-\rho}{1+\rho}\right]^{\frac{1}{2}}, \frac{1-\rho}{1+\rho}\right\}\right),$$

□

C Proofs of section 4:

Proof of theorem 5: The Taylor expansions of $\check{A}_\alpha^h$ and $\check{A}_\alpha^{h-1}$ around A_α is

$$\check{A}_\alpha^k = A_\alpha^k + \sum_{j=0}^{k-1} A_\alpha^j(\check{A}_\alpha - A_\alpha)A_\alpha^{k-1-j} + O_p(T^{-1}),\ k = h, h-1.$$

Then, using that $\sum_{j=0}^{h-2} A_\alpha^j(\check{A}_\alpha - A_\alpha)A_\alpha^{h-2-j} = \sum_{j=1}^{h-1} A_\alpha^{j-1}(\check{A}_\alpha - A_\alpha)A_\alpha^{h-1-j}$,

$$\begin{aligned}\check{w}_{T+h} - w_{T+h} &= -L_1 + L_2 + e'_{p+2}A_\alpha^0(\check{A}_\alpha - A_\alpha)A_\alpha^{h-1}Y_T \\ &\quad + e'_{p+2}\sum_{j=1}^{h-1} A_\alpha^{j-1}(A_\alpha - I_{p+2})(\check{A}_\alpha - A_\alpha)A_\alpha^{h-1-j}Y_T + O_p(T^{-1}). \qquad \text{(C.1)}\end{aligned}$$

Given that $(\check{A}_\alpha - A_\alpha) = e_{p+2}(\check{\varphi} - \varphi)'$, we can rewrite (C.1) as

$$\check{w}_{T+h} - w_{T+h} = (C'_{h,1} + C'_{h,2})Y_T - L_1 + L_2 + O_p(T^{-1}),$$

where

$$C'_{h,1} = e'_{p+2}A_\alpha^0 e_{p+2}(\check{\varphi} - \varphi)'A_\alpha^{h-1} \qquad \text{(C.2)}$$

$$C'_{h,2} = \sum_{j=1}^{h-1} e'_{p+2}A_\alpha^{j-1}(A_\alpha - I_{p+2})e_{p+2}(\check{\varphi} - \varphi)'A_\alpha^{h-1-j}. \qquad \text{(C.3)}$$

Taking expectation to the square of (C.1) we have

$$\begin{aligned}E(\check{w}_{T+h} - w_{T+h})^2 &= E(L_1 - L_2)^2 + E(C'_{h,1}Y_TY'_TC_{h,1}) \\ &\quad + E(C'_{h,2}Y_TY'_TC_{h,2}) + 2E(C'_{h,1}Y_TY'_TC_{h,2}) + o(T^{-1}). \qquad \text{(C.4)}\end{aligned}$$

The first term at the right side of (C.4) verifies that

$$\begin{aligned}L_1 - L_2 &= \sum_{k=0}^{h-1} e'_{p+2}A_\alpha^k U_{t+h-k,p+2} - \sum_{k=1}^{h-1} e'_{p+2}A_\alpha^k U_{t+h-k,p+2} \\ &= a_{T+h} + \sum_{k=1}^{h-1} e_{p+2}A_\alpha^{k-1}(A_\alpha - I_{p+2})e_{p+2}a_{T+h-k}.\end{aligned}$$

If we denote as $\psi_{h[\mathrm{AR}(p+1)]}$ to the h-th coefficient of $\varphi(B)^{-1}$ and $\psi_{h[\mathrm{ARIMA}(p+1,1,1)]}$ to the h-th coefficient of $\varphi(B)^{-1}(1-B)$, then

$$\begin{aligned}e'_{p+2}A_\alpha^{h-1}(A_\alpha - I_{p+2})e_{p+2} &= \psi_{h[\mathrm{AR}(p+1)]} - \psi_{h-1[\mathrm{AR}(p+1)]} \\ &= \psi_{h[\mathrm{ARIMA}(p+1,1,1)]} = e'_{p+2}A_1^k c_{p+2},\end{aligned}$$

and hence

$$L_h = L_1 - L_2 = \sum_{k=0}^{h-1} e'_{p+2}A_1^k c_{p+2}a_{T+h-k}. \qquad \text{(C.5)}$$

Then

$$E(L_h)^2 = \sigma^2 \sum_{k=0}^{h-1}(e_{p+2}A_1^h c_{p+2})^2.$$

It is also verified that

$$(e'_{p+2}A_1^h c_{p+2}) = (e'_p A_p^h e_p) + O(1-\rho) = (e'_p A_p^h e_p) + o(T^{-1}). \tag{C.6}$$

Applying this to $E(C'_{h,2}Y_T Y'_T C_{h,2})$ we obtain

$$\begin{aligned} E(C'_{h,2}Y_T Y'_T C_{h,2}) &= \sum_{j=0}^{h-2}\sum_{k=0}^{h-2}(e'_p A_p^j e_p)(e'_p A_p^k e_p) \\ &\quad \times \text{trace}\left\{E\left((\breve{\varphi}-\varphi)' A_\alpha^{h-1-j} Y_T Y'_T A_\alpha'^{\,h-1-k}(\breve{\varphi}-\varphi)\right)\right\}. \end{aligned}$$

Moreover, since the dependency between Y_T and $\breve{\varphi}$ is $O(T^{-\frac{3}{2}})$ (Kunitomo & Yamamoto 1985) and applying that

$$\text{MSE}(\breve{\varphi}) = \frac{\sigma^2}{T}\Gamma_y^{-1} + O(T^{-\frac{3}{2}}),$$

we obtain

$$\begin{aligned} E(C'_{h,2}Y_T Y'_T C_{h,2}) &= \frac{\sigma^2}{T}\sum_{j=0}^{h-2}\sum_{k=0}^{h-2}(e'_p A_p^j e_p)(e'_p A_p^k e_p) \\ &\quad \times \text{trace}\left\{A_\alpha^{h-1-j}\Gamma_y A_\alpha'^{\,h-1-k}\Gamma_y^{-1}\right\} + o(T^{-1}). \end{aligned} \tag{C.7}$$

Similarly it can be proven that

$$\begin{aligned} E(C'_{h,1}Y_T Y'_T C_{h,1}) &= \frac{\sigma^2}{T}(e'_p A_p^0 e_p)(e'_p A_p^0 e_p) \\ &\quad \times \text{trace}\left(A_\alpha^{h-1}\Gamma_y A_\alpha^{h-1}\Gamma_y^{-1}\right) + o(T^{-1}), \end{aligned} \tag{C.8}$$

where we have maintained the terms A_α^0 to ease the comparison with (C.7). Also

$$\begin{aligned} E(C'_{h,1}Y_T Y'_T C_{h,2}) &= \frac{\sigma^2}{T}\sum_{j=1}^{h-1}(e'_p A_p^j e_p)(e'_p A_p^0 e_p) \\ &\quad \times \text{trace}\left(A_\alpha^{h-1-j}\Gamma_y A_\alpha^{h-1}\Gamma_y^{-1}\right) + o(T^{-1}). \end{aligned} \tag{C.9}$$

Adding (C.8), (C.7) and two times (C.9) we obtain the second term of the right side of (4.8). Let us now proof (4.9):

$$\begin{aligned} E[(\breve{w}_{T+h} - w_{T+h})(\breve{w}_{T+k} - w_{T+k})] &= E(L_h L_k) + E(C'_{h,1}Y_T Y'_T C_{k,1}) + E(C'_{h,1}Y_T Y'_T C_{k,2}) \\ &\quad + E(C'_{h,2}Y_T Y'_T C_{k,1}) + E(C'_{h,2}Y_T Y'_T C_{k,2}) + o(T^{-1}), \end{aligned}$$

where $C_{h,1}$, $C_{h,2}$, $C_{k,1}$ and $C_{k,2}$ are as in (C.2) and (C.3) and L_h, L_k are as L_h in (C.5). Then

$$\begin{aligned} E(L_h L_k) &= E\left[\left(\sum_{j=0}^{h-1} e'_{p+2} A_1^j c_{p+2} a_{T+h-j}\right)\left(\sum_{i=0}^{k-1} e'_{p+2} A_1^i c_{p+2} a_{T+k-i}\right)\right] \\ &= \sigma^2 \sum_{j=0}^{h-1} (e'_{p+2} A_1^j c_{p+2})(e'_{p+2} A_1^{i+(k-h)} c_{p+2}). \end{aligned}$$

The remaining terms are solved as in the proof of (4.8). The values of s_0 needed in this theorem are more restrictive than in previous theorems in order to guarantee the aplicability of Yamamoto & Kunitomo (1985) results. □

Proof of theorem 6: The expectation of the square of (4.12) is

$$\begin{aligned} E[(w_{T+h} - \hat{w}_{T+h})^2] &= E(L_h^2) + E\left[e'_p(\hat{A}_p^h - A_p^h) W_T W_T' (\hat{A}_p^h - A_p^h)' e_p\right] \\ &\quad + E\left[\left(\sum_{j=h-1}^{\infty} \psi_j (1-\rho) z_{T-1-j}\right)^2\right] + E\left[\left(\sum_{j=0}^{h-2} \psi_j (1-\rho)\rho^{h-1-j} z_T\right)^2\right] \\ &\quad + 2E\left[e'_p(\hat{A}_p^h - A_p^h) W_T \sum_{j=h-1}^{\infty} \psi_j (1-\rho) z_{T-1-j}\right] \\ &\quad + 2E\left[e'_p(\hat{A}_p^h - A_p^h) W_T \sum_{j=0}^{h-2} \psi_j (1-\rho)\rho^{h-1-j} z_T\right] \\ &\quad + 2E\left[\left(\sum_{j=h-1}^{\infty} \psi_j (1-\rho) z_{T-1-j}\right)\left(\sum_{j=0}^{h-2} \psi_j (1-\rho)\rho^{h-1-j} z_T\right)\right]. \end{aligned}$$

The term L_h is the same than in (C.5), therefore

$$E(L_h)^2 = \sigma^2 \sum_{k=0}^{h-1} (e_{p+2} A_1^h c_{p+2})^2.$$

Applying (A.6) with $k = 1$

$$E\left[\left(\sum_{j=h-1}^{\infty} \psi_j (1-\rho) z_{T-1-j}\right)^2\right] = O\left(\frac{1-\rho}{1+\rho}\right) = o(T^{-1}).$$

Similarly

$$E\left[\left(\sum_{j=0}^{h-2} \psi_j (1-\rho)\rho^{h-1-j} z_T\right)^2\right] = O\left(\frac{1-\rho}{1+\rho}\right) = o(T^{-1}),$$

and, by Hölders' inequality

$$E\left[\left(\sum_{j=h-1}^{\infty}\psi_j(1-\rho)z_{T-1-j}\right)\left(\sum_{j=0}^{h-2}\psi_j(1-\rho)\rho^{h-1-j}z_T\right)\right] = O\left(\frac{1-\rho}{1+\rho}\right) = o(T^{-1}).$$

We will use a Taylor expansion of $\hat{A}_p$ around A_p. The magnitude of the remainder term is determined by the root-T consistency of $\hat{A}_p$

$$\begin{aligned}\hat{A}_p^h &= A_p^h + \sum_{j=0}^{h-1} A_p^j(\hat{A}_p - A_p)A_p^{h-1-j} \\ &\quad + \sum_{j=1}^{h-1}\left(\sum_{k=0}^{j-1} A_p^k(\hat{A}_p - A_p)A_p^{j-1-k}\right) \times (\hat{A}_p - A_p)A_p^{h-i-j} + O_p(T^{-\frac{3}{2}}).\end{aligned}$$

Therefore

$$e_p'(\hat{A}_p^h - A_p^h)W_T = (B_{h,1}' + B_{h,2}')W_T, \tag{C.10}$$

where

$$\begin{aligned}B_{1h,p}' &= e_p' \sum_{j=0}^{h-1} A_p^j(\hat{A}_p - A_p)A_p^{h-1-j} \\ B_{2h,p}' &= e_p' \sum_{j=1}^{h-1}\left(\sum_{k=0}^{j-1} A_p^k(\hat{A}_p - A_p)A_p^{j-1-k}\right) \times (\hat{A}_p - A_p)A_p^{h-i-j}.\end{aligned}$$

Taking expectation to the square of (C.10)

$$\begin{aligned}E[e_p'(\hat{A}_p^h - A_p^h)W_T W_T'(\hat{A}_p^h - A_p^h)'e_p] &= E(B_{h,1}'W_T W_T'B_{h,1}) + E(B_{h,2}'W_T W_T'B_{h,2}) \\ &\quad +2E(B_{h,1}'W_T W_T'B_{h,2}).\end{aligned}$$

Applying lemma A.3 we have

$$E(\|\hat{A}_p - A_p\|^k) \leq E(\|\hat{A}_p - A_p\|)^2)^{\frac{k}{2}} = O\left(\left[\frac{1-\rho}{1+\rho}\right]^k 2\right) = o(T^{-\frac{k}{2}}). \tag{C.11}$$

Consequently, applying Hölders' inequality $E(B_{h,2}'W_T W_T'B_{h,2}) = o(T^{-2})$, $E(B_{h,1}'W_T W_T'B_{h,2}) = o(T^{-\frac{3}{2}})$. It also holds that

$$\begin{aligned}E[e_p'(\hat{A}_p^h - A_p^h)W_T \sum_{j=h-1}^{\infty}\psi_j(1-\rho)z_{T-1-j}] &= o(T^{-1}), \\ E[e_p'(\hat{A}_p^h - A_p^h)W_T \sum_{j=0}^{h-2}\psi_j(1-\rho)\rho^{h-1-j}z_T] &= o(T^{-1}).\end{aligned}$$

Moreover

$$E(B'_{h,1}W_TW'_TB_{h,1}) = \sum_{j=0}^{h-1}\sum_{k=0}^{h-1}(e'_pA_p^je_p)(e'_pA_p^ke_p) \times \text{trace}\left\{E\left((\hat{\phi}-\phi)(\hat{\phi}-\phi)'A_p^{h-1-j}W_TW'_TA_p'^{h-1-k}\right)\right\}.$$

Applying theorem 4 and Hölders' inequality

$$E\left((\hat{\phi}-\phi)(\hat{\phi}-\phi)'A_p^{h-1-j}W_TW'_TA_p'^{h-1-k}\right) = E\left((\hat{\phi}_{|p}-\phi)(\hat{\phi}_{|p}-\phi)'A_p^{h-1-j}W_TW'_TA_p'^{h-1-k}\right) + o(T^{-1}).$$

Now, applying that the dependency between $\hat{\phi}_{|p}$ and W_T is $O(T^{-\frac{3}{2}})$ (Kunitomo & Yamamoto (1985)) and that $\text{MSE}(\hat{\phi}_{|p}) = \sigma^2/(T-1)\Gamma_w^{-1} + O(T^{-3/2})$,

$$E\left((\hat{\phi}-\phi)(\hat{\phi}-\phi)'A_p^{h-1-j}W_TW'_TA_p'^{h-1-k}\right) = \frac{\sigma^2}{T-1}(A_p^{h-1-j}\Gamma_wA_p'^{h-1-k}\Gamma_w^{-1}).$$

We complete the proof by applying that

$$\frac{\sigma^2}{T-1} = \frac{\sigma^2}{T} + O(T^{-2}).$$

□

The proof of (4.14) is a direct application of this proof and the proof of (4.9).The values of s_0 needed in this theorem are more restrictive than in previous ones in order to guarantee the aplicability of Yamamoto & Kunitomo (1985) results. □

D Proofs of section 5:

Proof of lemma 1: Let us decompose Y_t as $Y_t = (\tilde{Y}'_t,\ 0)' + \boldsymbol{\mu}$, where $\boldsymbol{\mu} = (\mu, \mu, ..., \mu, 1)'$. Since $\alpha = \mu - \sum_{i=1}^{p+1}\varphi_i\mu$ it is verified that $A_\alpha^i\boldsymbol{\mu}A_\alpha'^j = \bar{\boldsymbol{\mu}}$, where $\bar{\boldsymbol{\mu}} = \boldsymbol{\mu}\boldsymbol{\mu}'$. Then $A_\alpha^i\Gamma_yA_\alpha'^j = A_\alpha^i\Gamma_{\tilde{y}}^*A_\alpha'^j + \bar{\boldsymbol{\mu}}$, where $\Gamma_{\tilde{y}}^*$ is a $(p+2)\times(p+2)$ matrix with $\Gamma_{\tilde{y}}$ in the first $(p+1)\times(p+1)$ submatrix and zeroes elsewhere. Besides, the covariance matrix Γ_y has the following block structure

$$\Gamma_y = \begin{pmatrix} \Gamma_o & \boldsymbol{\mu}_o \\ \boldsymbol{\mu}'_o & 1 \end{pmatrix},$$

where $\Gamma_o = E(Y_{ot}Y_{ot}')$, with $Y_{ot} = (y_t, y_{t-1}, ..., y_{t-p})'$ and $\boldsymbol{\mu}_o = E(Y_{ot})$. Using the properties of the inverses of block matrices we can partition Γ_y^{-1} as

$$\Gamma_y^{-1} = \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix},$$

where $B_{11} = [\Gamma_o - \boldsymbol{\mu}_o\boldsymbol{\mu}_o']^{-1} = \Gamma_{\tilde{y}}^{-1}$. Then it is easily verified that $\text{trace}(A_\alpha^i \Gamma_{\tilde{y}}^* A_\alpha'^j \Gamma_y^{-1}) = \text{trace}(A_o^i \Gamma_{\tilde{y}} A_o'^j \Gamma_{\tilde{y}}^{-1})$ and then

$$\text{trace}(A_\alpha^i \Gamma_y A_\alpha'^j \Gamma_y^{-1}) = \text{trace}(A_o^i \Gamma_{\tilde{y}} A_o'^j \Gamma_{\tilde{y}}^{-1}) + \text{trace}(\boldsymbol{\mu}\Gamma_y^{-1}).$$

Given that $\text{trace}(\bar{\boldsymbol{\mu}}\Gamma_y^{-1}) = \boldsymbol{\mu}'\Gamma_y^{-1}\boldsymbol{\mu}$ and applying Searle (1984, pag 258) it can be seen that

$$\boldsymbol{\mu}'\Gamma_y^{-1}\boldsymbol{\mu} = 1 - \mid \Gamma_y - \boldsymbol{\mu}\boldsymbol{\mu}' \mid / \mid \Gamma_y \mid = 1,$$

since the last column and row of $\Gamma_y - \boldsymbol{\mu}\boldsymbol{\mu}'$ are zeroes and Γ_y is invertible. □

Proof of lemma 2: In this proof we will follow the arguments of Box & Tiao (1977), where they decompose a nearly nonstationary vector process into two parts, the first part following a stationary process and the second one approaching nonstationarity. Let C be the following nonsingular matrix

$$C = \begin{pmatrix} 1 & -\rho & 0 & \cdots & 0 & 0 \\ 0 & 1 & -\rho & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & -\rho \\ 1 & -\phi_1 & -\phi_2 & \cdots & -\phi_{p-1} & -\phi_p \end{pmatrix}.$$

Then

$$CA_oC^{-1} = \begin{pmatrix} A_p & 0 \\ 0 & \rho \end{pmatrix} = D.$$

The nonsingularity of C can be seen, for instance in Pearl (1973, p. 156). It is easily verified that

$$(CA_oC^{-1})^i = CA_o^iC^{-1} = \begin{pmatrix} A_p^i & 0 \\ 0 & \rho^i \end{pmatrix} = D^i. \tag{D.1}$$

Let λ_k be an eigenvalue of the matrix $Q = \Gamma_{\tilde{y}}^{-1} A_o^i \Gamma_{\tilde{y}} A_o'^j$. Then λ_k is a root of the determinantal polynomial

$$|Q - \lambda I| = |A_o^i \Gamma_{\tilde{y}} A_o'^j - \lambda\Gamma_{\tilde{y}}| = 0.$$

Using (D.1) this expresion is equivalent to

$$|D^i \Gamma_C D'^j - \lambda \Gamma_C| = 0, \tag{D.2}$$

where $\Gamma_C = C\Gamma_{\tilde{y}}C'$. This matrix Γ_C can be considered as the covariance matrix of the transformed series $Z_t = CY_t$, where $Z_t = (z_{1,t}, z_{1,t-1}, ..., z_{1,t-p+1}, z_{2,t})'$ and

$$Z_t = DZ_{t-1} + a_t c_{p+1}, \tag{D.3}$$

with $c_{p+1} = (1, 0, ..., 0, 1)'$. Therefore the first $p \times p$ submatrix of Γ_C is the covariance matrix of a process following the coefficient matrix A_p and noise a_t, namely the matrix $\Gamma_{w|p}$. Denoting as V_{12}, V_{21} and V_{22} the other submatrices of this partitioning we can rewrite (D.2) as

$$\begin{vmatrix} (A_p^i \Gamma_{w|p} A_p'^j - \lambda \Gamma_{w|p}) & (A_p^i V_{12} \rho^j - \lambda V_{12}) \\ (\rho^i V_{21} A_p'^j - \lambda V_{21}) & (\rho^{i+j} V_{22} - \lambda V_{22}) \end{vmatrix} = 0,$$

which can also be expressed as

$$\begin{vmatrix} (A_p^i \Gamma_{w|p} A_p'^j - \lambda \Gamma_{w|p}) & (A_p^i V_{12} \rho^j - \lambda V_{12})(V_{22}^{-\frac{1}{2}}) \\ (\rho^i V_{21} A_p'^j - \lambda V_{21})(V_{22}^{-\frac{1}{2}}) & (\rho^{i+j} - \lambda) \end{vmatrix} = 0.$$

Following (D.3) The term V_{22} is the variance of an AR(1) process with coefficient ρ and therefore $V_{22}^{-1} = O(1-\rho)$. Hence, following the rule to evaluate the determinant of a partitioned matrix (see, for instance, Searle 1984)

$$|Q - \lambda I| = \left|A_p^i \Gamma_{w|p} A_p'^j - \lambda \Gamma_{w|p}\right| \left(\rho^{i+j} + O(1-\rho) - \lambda\right) = 0.$$

Since the trace of a matrix equals the sums of the eigenvalues the lemma holds. □

Proof of theorem 8: If $\beta = 1 - v$, with $v \geq 0$, we can not substitute the terms $(e'_{p+2} A_1^v e_{p+2})$ by $(e'_p A_p^v e_p)$, as made in (C.6). This makes both PMSE increase their distance. Also, and more important, most of the arguments for the proofs of theorem 6 lie on the assumption that the term $(1-\rho)(1+\rho)^{-1}$ is $o(T^{-1})$, which is true only if $\beta > 1$. This introduces an error term in expresions 4.13 and 4.14 of magnitude $O(T^{-1+v})$. Lemma 2 also need that $\beta < 1$ to hold. □